Bond University

DOCTORAL THESIS

Multiservice Traffic Allocation in LEO Satellite Communications.

Septiawan, Reza

Award date:
2004

MULTISERVICE TRAFFIC ALLOCATION IN LEO SATELLITE COMMUNICATIONS

by

Reza Septiawan

Submitted to the Faculty of Information Technology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

BOND UNIVERSITY

July 2004

Signature of Author .....
Faculty of Information Technology
July 2004

Certified by .....
Stephen Sugden
Dr, Associate Professor
Thesis Supervisor

Accepted by .....
Chairperson, Research Committee on Graduate Students

MULTISERVICE TRAFFIC ALLOCATION IN LEO SATELLITE COMMUNICATIONS

by

Reza Septiawan

Submitted to the Faculty of Information Technology
on July 2004, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Satellite communication promises potential methods for providing global communication. In particular, by the development of a Low Earth Orbital (LEO) satellite constellation, both global coverage and broadband communication will be accessible. Problems arise in situations where various traffic types in broadband communication require different levels of quality of service (QoS). Traffic control is required to make sure that each traffic demand may receive the expected QoS. Another problem is that the dynamic topology of a LEO satellite network requires a traffic allocation control, which is able to allocate traffic demand into the Inter Satellite Links (ISLs) between LEO satellites.

In this thesis, traffic allocation strategy in a dynamic LEO satellite communication network is studied and analyzed. The delivery of Quality of Service (QoS) is an important objective. Traffic allocation control is performed in the LEO satellite constellation to provide a near optimal utilization of these ISLs. An alternative solution is proposed in this research, in which a combination of two algorithms will be used to allocate traffic in this dynamic satellite network. The first algorithm allocates traffic during small time intervals, based on an assumption that the topology is unchanged during these intervals. The second algorithm allocates traffic after topology updating has been accomplished. Traffic allocation respects some constraints including QoS (due to multiservice requirements), capacity constraints, traffic distribution, and availability constraints. Both theoretical and empirical studies have been undertaken to examine the performance of the proposed algorithm, denoted GALPEDA (Genetic Algorithm Linear Programming and Extended Dijkstra Algorithm). The proposed algorithm provides privileges to a class of high priority traffic, including benefits for traffic allocation of multiclass traffic in LEO satellite communication. It provides a novel traffic allocation mechanism to cope with the dynamic topology of a LEO satellite; moreover this algorithm distributes multiservice traffic evenly over the network. Simulations results are provided.

Thesis Supervisor: Stephen Sugden
Title: Dr, Associate Professor

Contents

List of Figures

List of Tables

Acknowledgements

Pursuing a doctoral degree is a long journey. It is not undertaken alone and consequently, there are many people who I would like to thank for their contribution.

First, I would like to thank my parents for all the love, support and encouragement they have provided me. They have made many sacrifices to provide the wonderful opportunities that I have had. I am most grateful to them.

Graham McMahon and Stephen Sugden have been fantastic supervisors. They continually impressed me with their wisdom and many helpful discussions during my study and research at Bond University. I wish to thank them for always being available when I needed their advice. Moreover, they were excellent teachers especially, when it came to casino games and tennis. During the early period of my research, the Bond Algorithm Group helped me to solve some problems in relation to understanding different aspects of algorithms.

Thanks to Les Berry from RMIT for his significant advice in teletraffic engineering, and to Zheng Da Wu for his advice on networking issues. Marcus Randall deserves thanks for sharing his experiences in Simulated Annealing, and Elliot Tonkes for helping me to solve mathematical equations. I wish to thank James Montgomery for giving me the opportunity to continue his research on 'gennet', which I have used as a starting point for my own research, and for proofreading my work at the end of my research period.

Thanks also, to Margareth DeMestre and Neville DeMestre for their advice. Thanks to Jessica Syme for editing my thesis. Clarence Tan provided advice and challenging questions regarding mobile communication, which sharpened my research. I wish to thank him also for providing the opportunity to learn about neural networking. Additionally, he gave me the opportunity to learn badminton from him and Paddy Krishnan.

I thank Ron Luken from Reach-Telstra, Sydney for giving me the opportunity to learn more about their satellite sites during my research in Sydney. Thanks must go to - Tiok Woo Teo who provided much help with all kinds of technical and non-technical problems - Cyrille Clipet for guiding me through the difficulties of setting up network simulator (ns) and using Linux. I also wish to thank my fellow research students and friends who have 'come and gone' during my long stay in the research room, room 5323 IT School. I am most grateful to all the IT School staff who helped me during my research at Bond.

Special thanks to the ARC (Australian Research Council) large and small grant, which supported our research, and gave me the opportunity to purchase a very useful computer. Thanks also, to BPPT (the Agency for Assessment and Application Technology of Indonesia) which provided the opportunity for me to study overseas, and jointly with the IT School at the end of my study period, provided my allowance and scholarship permitting the continuation of my work.

Finally, I would like to thank my wife and her family, who supported me and accompanied me throughout this long journey of pursuing my doctorate. Her patience and support are highly valued.

Abbreviations and Acronyms

ABR	Available Bit Rate
AMPS	Advanced Mobile Phone System
ARMA	Auto Regressive Moving Average
ATDMA	Asynchronous TDMA
ATM	Asynchronous Transfer Mode
BER	Bit Error Rate
Bpsat	Bent Pipe Satellite
CBR	Constant Bit Rate
CDMA	Code Division Multiple Access
DAMA	Demand Assignment Multiple Access
D-AMPS	Digital AMPS

DCR	Dynamic Control Routing
DCS	Digital Communication System
DDS	Digital Data Service
DNHR	Dynamic Non Hierarchical Routing
DSCP	Differentiated Services Codepoint
EDA	Extended Dijkstra shortest path Algorithm
EDGE	Enhanced Data Rate for GSM Evolution
EIR	Equipment Identity Register
ERP	Effective Radiated Power
ES	Earth Station
ETSI	European Telecommunication Standards Institute
EURESCOM	European Institute for Research and Strategic Studies in
FDMA	Frequency Division Multiple Access
FHRP	Footprint Handover Rerouting Protocol
FIFO	First In First Out

FITCE	Federation of Telecommunication Engineers of the European
GA	Genetic Algorithm
GALPEDA	Genetic Algorithm Linear Programming Extended Dijkstra
GEO	Geostationary Earth Orbit
GPRS	General Packet Radio Service
GSM	Groupe Speciale Mobile, Global System for Mobile communication
GSO	Geosynchronous Orbit
GW	Gateway
HEO	Highly Elliptical Orbit
HLR	Home Location Register
HSCSD	High Speed Circuit Switched Data
HVS	Human Visual System
IETF	Internet Engineering Task Force
ILP	Integer Linear Programming
IP	Internet Protocol

IPv4	Internet Protocol version 4
IPv6	Internet Protocol version 6
ISL	Inter Satellite Link
ISO	International Standards Organization
ITU	International Telecommunication Union
IWF	Inter Working Function
JDC	Japanese Digital Cellular
LAN	Local Area Network
LEO	Low Earth Orbit
LHS	Left Hand Side
LLC	Logical Link Control
LOS	Line of Sight
LP	Linear Programming
LUI	Last Useful Instant
MAC	Medium Access Control

MCNSP	Minimum Cost Network Synthesis Problem
MCR	Minimum Cell Rate
MEO	Medium Earth Orbit
MFA	Mean Field Annealing
MMPP	Markov Modulated Poisson Process
MSC	Mobile Switching Centre
MT	Mobile Terminal
ND	Neighbours Discovery
NOCC	Network Operations and Control Centre
NORAD	North American Aerospace Defence Command
OD	Origin Destination
OSI	Open System Interconnection
PASTA	Poisson Arrivals See Time Average
PDC	Personal Digital Cellular
PCN	Personal Communications Network

PLMN	Public Land Mobile Network
PSK	Phase Shift Keying
PSTN	Public Switched Telephone Network
QoS	Quality of Service
RAAN	Right Ascension of Ascending Node
RF	Radio Frequency
RHS	Right Hand Side
RLP	Radio Link Protocol
RRAA	Random Reservation Adaptive Assignment
RSVP	Resource Reservation Protocol
SDMA	Space Division Multiple Access
SGP	Simplified General Perturbation
SGP	Simplified General Perturbation
SIU	Satellite network Interface Unit
Swsat	Intelligent Switching Satellite

TACS	Total Access Communication System
TCP	Transmission Control Protocol
TCP/IP	Transmission Control Protocol/Internet Protocol
TDMA	Time Division Multiple Access
TINA-C	Telecommunication Information Networking Architecture Consortium
ToS	Type of Service
TSP	Traveling Salesman Problem
UBR	Unspecified Bit Rate
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunication System
VBR	Variable Bit Rate
VBR-NRT	VBR Non Real Time
VBR-RT	VBR Real Time
VLR	Visitor Location Register
Xcsat	Cross Connect Satellit

Chapter 1

INTRODUCTION

1.1 Introduction

The requirement for more bandwidth has placed considerable pressure on wireless network operators, and put stress on the need for efficient use of limited spectrum allocations. This is especially true in urban areas. Currently, due to the high user demand of broadband communication, the amount of bandwidth that a system can allocate to individual subscribers is restricted. As a result, wireless network operators are investigating new techniques that will overcome the bandwidth limitation, and allow an inexpensive addition of system capacity whilst providing a high bandwidth capability.

The growing number of users interested in multimedia communication has also increased the demand for more bandwidth. Transporting multimedia communication over wireless environments needs extra consideration because wireless environments are not as reliable as wired networks. Multimedia applications such as voice and video suffer from high variance of bandwidth and Bit Error Rate (BER) in wireless environments. Such real time applications rely on the appropriate delivery of data. In wireless environments, these applications need to cope with 'physical layer' errors induced by path-loss, fading, channel interference and shadowing.

Another aspect, which was introduced at the end of the 20^{th} century, is the significant need for mobility. There is a need to be globally accessible by using a portable mobile terminal.

Both multimedia data communications and mobile telephony are growing simultaneously. Current systems make use of both technologies and a new era for communications has become

a reality: mobile multimedia communication.

Future demands for mobile multimedia communication systems will be characterized by heterogeneity of broadband services, which are to be supplied in indoor and outdoor environments, simultaneously, with varying degrees of mobility. In particular, the increase use of MMS (Multimedia Message Service), IM (Instant Messaging) etc. and the increased number of mobile subscribers means more bandwidth is required. A current report from In-Stat/MDR [1] stated that in the next five years, there will be a slowing down in mobile subscriber growth. Yet, there will be more than 931 million new subscribers by the year 2007. This means that there will be a total wireless population of more than two billion subscribers worldwide and an annual subscriber growth of 186 million on average (figure 1.1.1).

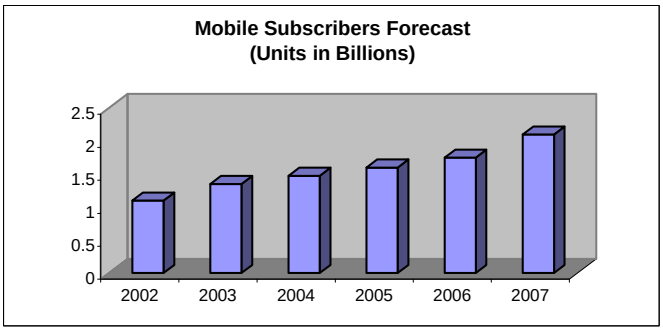

Figure 1.1.1: Mobile subscribers in 2002 and the forecast number from 2003 to 2007

In-stat/MDR reported that in the Asian Pacific region, an explosive growth in messaging volumes is forecasted throughout the period from 2003 till 2007 [2]; see figure 1.1.2.

Most of these new types of activities will require high bandwidth. A complementary communication medium to terrestrial mobile communication is essential to provide global communication. If we attempt to provide a broadband service to customers, then the first alternative solution is expanding the bandwidth. However, this is only a part of the solution. It is possible to use the same amount of bandwidth more effectively with good traffic control. Optimal traffic control makes it possible to route calls directly to mobile users with their required QoS, regardless of their locations. This approach will improve the quality of service for broadband communication and will help to solve the problem.

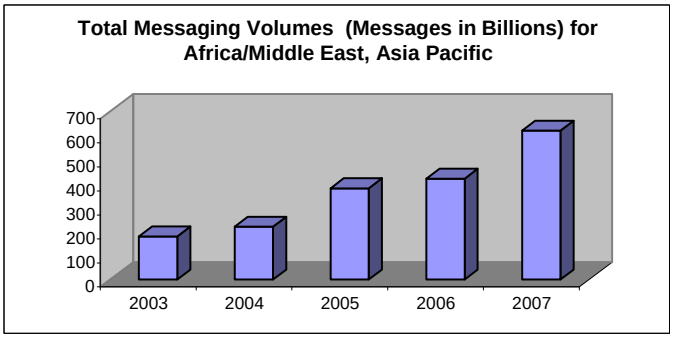

Figure 1.1.2: Forecast total messaging volumes from 2003 to 2007

1.2 Traffic Management

The previously mentioned new multimedia applications have driven researchers to explore means of traffic control that can cope with the increase number of these applications. An important goal is to implement a routing algorithm that can optimally allocate traffic in the network. In addition to the increased number of multimedia applications with various Quality of Service (QoS) requirements, it is essential to have a routing algorithm that is capable of providing a guarantee to the user of meeting their expected QoS requirements, or their required various level of QoS (multi-service level of QoS).

In this thesis, a multi-service routing algorithm is proposed, which will give some preference to demand with high QoS. The novelty of our proposed multi-service routing algorithm is that it will consider the remaining bandwidth on each satellite link when allocating different traffic class types, so that an evenly distributed traffic load can be achieved.

1.3 Satellite Communication Development

The introduction of mobile multimedia communication and the Internet has removed the distance barrier between users around the world. Satellite communication can complement the already existing terrestrial network and improve the coverage of the network. Satellite communication and the existing terrestrial network can provide global coverage for users. The most used satellite communication, Geo-stationary Earth Orbital (GEO) satellite communication,

has been used over many years as a significant alternative communication tool in rural areas. GEO satellite communication is used to provide a global connection for users. However, the high orbital positions of these satellites lead to some drawbacks in their function as a communication medium.

One significant weakness of GEO satellite communication is that of long delay. This is especially important in providing connections for a delay sensitive application, which is essential in multimedia communication. Therefore, lower orbit satellites, LEO satellites have been launched in order to reduce the drawbacks of GEO systems. The LEO satellites provide a shorter delay transmission, and due to the lower orbital position, less transmitting power is required. A smaller receiver antenna can be used to receive the signal. On the other hand, the lower altitude of this satellite means a faster moving satellite, which will reduce the satellite visibility period. LEO satellites are only visible for a few minutes from the user's position on earth. Therefore, global communication can only work when a large number of satellites in the LEO system are operational. In this thesis, we, propose a technique to cope with the complexity of allocating the incoming traffic into LEO satellite links. We introduce a combination algorithm GALPEDA, which will cope with the dynamic property of the LEO satellite network.

1.4 Contributions of This Work

The work in this thesis has made contributions to broadband satellite communication in the following ways. We propose a combination algorithm to cope with the dynamic topology of LEO satellite communication which combines the implementation of Genetic Algorithm and Linear Programming (GALP) with an extended Dijkstra shortest path Algorithm (EDA). Our proposed algorithm can give privileges to high priority class traffic and, in addition, provides a more distributed traffic load over the network [3].

In order to deal with the dynamic property of LEO satellite networks, a combination of two different algorithms is applied to solve the traffic allocation problem. The first algorithm, GALP, allocates traffic periodically to tackle the changing LEO satellite topology. The second algorithm, EDA, allocates traffic between these periodic updates. The time interval between periodic updates is partly determined by the current traffic load. If the traffic load of the LEO

satellite network is higher than a threshold value, then periodic updates will be performed more frequently. In order to be able to use the same satellite for the duration of the time interval, the time intervals must be smaller than the visible time of the LEO satellite from a user on earth. This approach allows us to simplify the handover problems.

GALP solves traffic allocation problem of the LEO satellite network, globally, ensuring an evenly distributed load of traffic around the globe. EDA solves the traffic allocation problem of the network locally, handling changes at a local level. The combination provides an evenly distributed load of traffic and an efficient traffic allocation process.

In both algorithms, we introduce a parameter which is attended to give privileges for high priority traffic. The privileges given are related to the remaining bandwidth in the Inter Satellite Link. If the remaining bandwidth in an ISL is low, then a low priority traffic will have more chance to be rerouted into a longer path than a high priority traffic.

GALP in this thesis has two characteristics, which are preferable in this implementation of this algorithm into LEO satellite network. Firstly, we introduced the mutation character of Genetic algorithm. This mutation property helps us to solve the problem of local optima. When a mutation should occur according to a certain mutation probability, then we choose one solution in the previous population and use it as one of the parents (we use this solution without updating their new satellite positions). In this GA version, instead of having only two chromosomes as parents, the third 'parent' is introduced. The 'third' parent originates from the previous time interval's feasible solution. We update this solution according to their new LEO satellite positions (we use the previous solution with updating of their new satellite positions). This introduction of the third 'parent' has improved the traffic allocation processing time. On the other hand, this third 'parent' helps us to provide possible handover options into GALP. Since we can predict the positions of the satellite in the new time interval, then we will have a set of connections, which have to be handed over. In case of this handover, the middle path remains the same and we add an extra path from the old source to a new source and from the old destination to a new destination. When a new call arrives inside the time intervals, EDA allocates a path from source to destination with an objective to give privilege to high priority traffic, and performs a more distributed traffic. In the case that there is no available link capacity to be given, low priority traffic is diverted into a longer path than high priority

traffic. Due to our algorithm low priority traffic will have a longer path than high priority traffic but a lighter loaded channel.

1.5 Thesis Outline

In the following chapters, a literature review is presented and background research of the topic is broadly described. Both terrestrial wired and wireless technology and their related advantages and disadvantages will be described.

In Chapter 4, LEO satellite communication as a complement of the existing terrestrial network will be outlined. Beginning with the available types of LEO satellite topology, we outline the signal processing in this satellite network. Chapter 5 gives the problem formulation, which represents criterion for optimization and constraints. Chapter 6 analyses the algorithms, which are used in our proposed multiservice routing algorithm. The individual algorithms will be described.

In Chapter 7, the proposed multiservice routing algorithm, GALPEDA is discussed. The simulation that we conducted to test our algorithm is introduced, assumptions have been made in order to reduce the complexity. In Chapter 8, the simulation results for our algorithm are evaluated. The graphical and tabular results are charted in the appendices. Finally, we present the thesis contributions, conclusions and future work plans in Chapter 9.

Chapter 2

TRAFFIC MODELLING, SATELLITE CONSTELLATION, AND TRAFFIC ALLOCATION

2.1 Introduction

Initiating broadband services into the mobile communication network requires the development of strategies to allocate this multimedia traffic in the wireless network. Different types of traffic require different types of QoS. Some applications need to have a secure connection between their origin and destination, while some other applications require developments in transmission rate. Multimedia traffic has also changed the traditional traffic characteristics [4–6]. Traffic in modern telecommunication networks is driven by the heterogeneous mix of traffic classes, which ranges from traditional telephone calls to video and data services.

Satellite communication offers a solution to deliver these heterogeneous traffic classes over the globe. Different types of satellite communication have been studied and implemented, which resulted in a design of broadband satellite systems that can provide high data rate transmission (1Mbps and above). As stated in [7], satellites can play an important role to deliver broadband services including a global internet by providing high-speed data transmission through a high-bandwidth capacity channel. The primary issue of delivering this service is: how can the QoS

be improved from the current best-effort service, and provide high-speed data access?

In this case, a routing algorithm which is priority sensitive (QoS-sensitive) is necessary to allocate this multi class traffic as given in [4, 6, 8–10]. In the first two papers QoS routing has been studied in terms of the processing complexity of determining QoS paths in link state based routing architectures. In order to reduce the processing overhead, caching mechanisms are used in [8]. Authors in [9] showed that network information inaccuracy can have a major impact on the complexity of determining the path selection process.

In this chapter, some research related to the topic of this thesis is discussed. Since network information accuracy has a major impact on the complexity of the path selection process, work related to traffic model and satellite networks will be reviewed; followed by some work related to traffic allocation either in terrestrial- or satellite-networks.

2.2 Traffic Model

In traditional telecommunication networks, the Poisson traffic model is used to model voice traffic [11]. In this model, the time intervals between two arrivals are exponentially distributed. There are no correlation between the arrivals; that is, each arrival is independent of other arrivals.

Currently, there is no model available that can exactly model heterogeneous multiclass traffic. Most models are based on a queuing model, in which traffic is offered to a queue of a network, and various performance measures are calculated. The performance measurements are the queue length, server utilization, waiting times and loss of traffic. Since modern telecommunication networks carry broadband traffic, the current traffic is characterized by its 'burstiness'. 'Bursty' traffic is generated at an uneven rate and it may have a strong correlation between arrivals. Markov based traffic models are often used to represent this 'bursty' traffic. 'Bursty' traffic occurs because of the need to transfer a high demand for traffic in a very short interval. This is subsequently followed by a long interval with a low demand for traffic. If we reserve some amount of transmission capacity at the peak rate, it would lead to a low average deployment of network resources. It is also concluded in the paper by Ake Arvidsson et al. [11] that it is hard to find a model which is universally acceptable and adaptable to the multiplicity of

different traffic types or classes. Other work by Jagerman et al. [12] surveyed some teletraffic models, addressing both theoretical and computational aspects. Some constraints in this work are heterogeneous traffic, 'bursty' traffic, instability of the network (especially in the wireless environment), Quality of Service requirements and the nature of video traffic i.e. that successive frames within a video scene vary little, but if the scene changes it can cause abrupt changes in frame bit rate.

T. Janevski [13] provides very useful information on a well-defined traffic theory. Basically telecommunication networks are divided into circuit switched and packet switched networks. Furthermore, based on their types of traffic we can classify networks into homogeneous and heterogeneous networks. If we consider the type of access network then we categorize telecommunications into wired (fixed) access networks and wireless (mobile) access networks. The telecommunication network is then a combination of these three different categorizations. In our case, we are focusing on wireless networks with heterogeneous traffic. In 1992, Bae et al. [14] analyzed multiple heterogeneous arrival streams and modeled the traffic following Markov Modulated Arrival Processes. The authors studied the behavior of loss probabilities due to the burstiness of traffic streams. They found that an increase in the burstiness of one stream results in an increase of packet loss probabilities of the stream itself and of the other streams, which are coming together. Since the current and future traffic in communication systems has self-similar properties in nature, in [15], the authors tried to estimate these self-similar properties by using two models of MMPP. In [16,17], real data traffic and mobility data has been used to perform a traffic model PCS. While in [18,19] they performed a real data measurement. The first paper measured backbone traffic variability, while the second paper measured trace data and modeled it into a synthetic traffic model, which is supposed to represent the 3G (UMTS) traffic. Shoji Kasahara [20] has studied the prediction of loss probability for finite queues by using the Markovian approach to self-similar traffic. Since heterogeneous traffic (such as Internet traffic) shows a self-similar nature, the author tried to model this in terms of MMPP. Other work has been conducted in relation to the mobility and traffic model in [21]. Those authors constructed a new analytical model that characterizes mobile user behavior and the resultant traffic patterns.

Studies of the performance of traffic models can be done in two ways [22]: either as part of a

theoretical test or as part of an empirical (statistical) test for example, to drive a discrete-event Monte Carlo simulation.

In our research, we focus on modeling traffic as heterogeneous traffic in either circuit or packet switched wireless networks. We used a model based on Monte Carlo simulation. It pseudo-randomly generates values for some variables in the LEO satellite constellation to simulate this constellation. In the future the traffic will have more self-similar type of traffic due to the burstiness of the packet traffic. However, many practitioners ignored this phenomenon because there are inadequate physical explanations for the observed self-similar nature of measured traffic from today's network; and there is a lack of studies on its impact on satellite network, protocol design and performance analysis [7]. Due to these reasons and the complexity reason in the simulation model, we consider only two models of traffic model, namely Poisson and MMPP.

2.3 Satellite Network

There are various research areas in satellite networks. The authors in [23] stated that satellite communications are categorized according to their functions and properties as fixed satellite systems, broadcast satellite systems, and mobile satellite systems as given in figure 2.3.1.

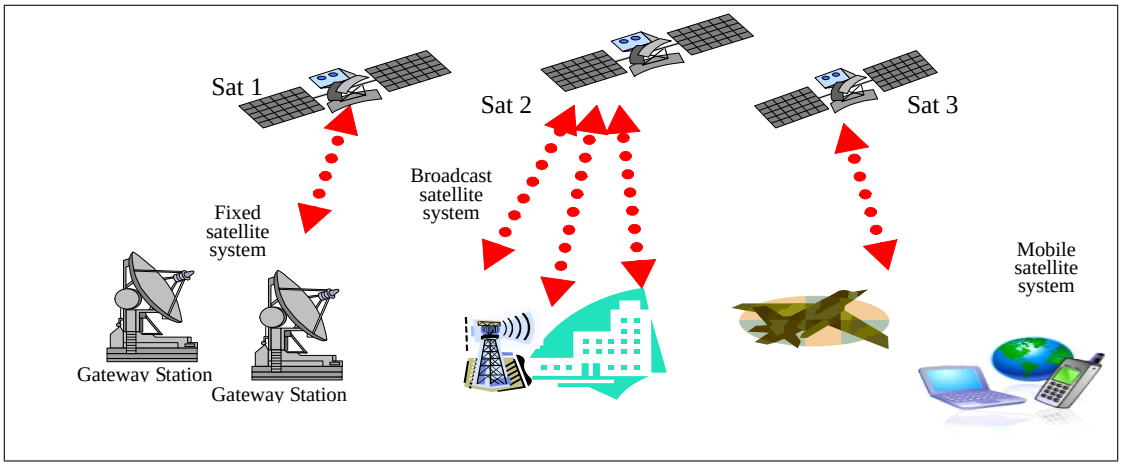

Figure 2.3.1: Various type of satellite system

Others include satellite research focusing on the study of the satellite vehicle itself; for

example, regarding the instruments of the satellite vehicles (battery, satellite coating material, launch rocket, etc.). Other studies focus more on the performance of satellite as a system; for example a LEO satellite communication system.

Some studies have been conducted to investigate the performance of LEO satellites, for example the simulation of LEO satellites in [24]. There are some characteristics which have a bigger effect in satellite than in terrestrial networks, for example the Doppler Effect. In LEO satellites, the Doppler shift due to satellite motion is greater than the Doppler shift in the cellular system [25, 26]. Others are looking to the performance of LEO satellites to deliver broadband services [27]. The authors in [28] studied the mobility aspect of LEO satellites. In addition, in [29] the effect of the time delay for frequency reuse and bandwidth utilization are investigated. Various losses occurred in satellite signals, which reflect in satellite network performance. According to Bekkers and Smits [30], shadow losses in LEO sstems typically follow a log-normal distribution which varies in the range 4dB to12dB in urban areas. Similar losses occur in LEO satellite systems but continuous changes in shadowed areas occurred since LEO satellites are moving

Other researchers focused on the architecture of the LEO satellite networks and their frequency-reuse. In the paper of Werner et al. [31], the authors provide us with system parameters of the already launched and planned satellite constellation. Various types of satellite topology will result in various performances of satellite networks. A variation of the Manhattan network is given in [32].

Our research will focus more on the performance of LEO satellite network as mobile satellite systems and their performances in respect to the dynamic characteristics of LEO satellites.

2.4 Traffic Allocation

After the first two components of our research topic, namely a traffic model and satellite network, the last component of our research is related to traffic allocation. Different traffic allocation strategies or different access methods could have a significant influence on the performance of the satellite constellation as a whole. There are some research in the area of traffic allocation schemes which are not implemented in LEO satellites. They could give us a better picture in

understanding these traffic allocation problems, since some similar traffic allocation schemes can be used to solve traffic allocation problems in LEO satellites.

In terrestrial communication networks there are interesting research in the traffic allocation, such as in [33]. The authors of this paper compared best effort services and reservation-capable services. They showed that across a wide range of bandwidth, reservation architecture reduced the total cost of a network to 10% of total cost in best effort architecture. Another type of routing algorithm, 'congestion dependent cost based algorithm' considered an exact and approximation approach of computation [34]. In this paper the author showed that by using an approximation approach, a reasonable performance of the routing algorithm could be obtained. Yang [35] investigated performances of a routing algorithm in different network configurations, with different configuration parameters such as CPU power, buffer size, router processing capability. Implementation of traffic allocation strategy in optical fiber networks has attracted some researchers to study QoS based routing algorithm (QoSRBF) [36], and random receiver algorithms for scheduling multicast traffic in Wavelength Division Multiplexing (WDM) [37–39]. They studied the provisioning for establishing a circuit from one end to another through the optical backbone. A traffic allocation algorithm in an IP network which schedules traffic from node to node and introduces a clue into the IP header is studied in [40]. In this approach, they add an extra 5 bits in the IP header to tell its downstream router where a good point to start for the IP lookup is. An IP-based routing algorithm is moreover studied in [41–43]. Previous papers consider ATM networks as their environment. Routing algorithms in a wireless environment are studied in [4, 44–46]. In a wireless environment, unique characteristics of wireless media ('bursty' channel errors and location dependent channel capacity and errors) should be introduced. A fair scheduling of delay and rate-sensitive packet flows can be performed using an adaptive routing algorithm, which is studied in [6, 47–49].

Kohei Shiomoto et al. [50] proposed a dynamic burst transfer time-slot-base network architecture which can cope with the bursty characteristics of current traffic load. This network architecture uses less space overhead since they use time division switching, in which switching is performed by exchanging time slots between input and output links and no packet header is required. A connection is established and released dynamically on a burst by burst basis. In order to accommodate bursty traffic, which occurs after sending routing information when

there is an insufficient time slot available, a buffer-pool and a buffer controller are used to keep the incoming bursty traffic.

M. B. Pursley et al. [51] proposed a Multimedia Least Resistance Routing, which use different link and path resistance metrics for different types of messages. The route chosen for a packet depends on the ability of the radio-waves along the route to receive and forward packets within the constraints required by different types of traffic. Through use of an adaptive routing algorithm, traffic allocation in a network could be updated according to the current condition of the network.

Different access methods and traffic allocation schemes in satellite communications are introduced in [52–56]. Research in LEO satellites has been done to investigate: the infrastructure and the possible handover mechanisms [57], [58–62], access protocol [63–66], and routing algorithm [32, 67–72]

Some papers assumed that the satellite networks are an IP based network, and studied the satellite performance. The authors in [73] and [74] introduced a new routing algorithm for multicast services. The first paper is implemented in LEO satellites, while the second paper should be implemented in the next generation of GEO satellites. The authors in [75], investigated a delay over a satellite network for voice over IP. An alternative solution for network layer integration of a hybrid IP network consisting of terrestrial and satellite IP networks has been proposed in [76]. The authors in this paper proposed a new Gateway protocol called Border Gateway Protocol-Satellite version (BGP-S), which enables automated discovery of paths that go through the satellite network.

Christopher Ward in 1995 [66] introduced a data link control protocol for LEO satellites. The author proposed a link layer protocol, which will be suited for LEO satellites and provides a reliable datagram service. The protocol is based on a negative acknowledgement (NAK) checkpoint protocol. Marc Emmelmann et al. in [63, 64] proposed an access protocol which guarantees a high ATM cell rate for a future multimedia ATM-based LEO satellite network. Another proposed Access protocol was given in [65], in which the authors proposed MAC protocol based on CDMA to support integrated services in LEO satellite systems. The protocol manages a flexible bandwidth allocated for DBR (Demanded Bit Rate) service by defining probabilities to access the channel. In the ATMSat project [77], the authors proposed an alternative MAC

protocol which can be used in a Ka-Band LEO satellite system with onboard ATM switching, signaling and resource management, optical ISLs and active intelligent antennas.

Some predictive resource allocation methods in LEO satellite network have been studied in recent years. The authors in [78] propose a predictive handoff management and admission control strategy for multimedia LEO satellite networks. An adaptive resource reservation protocol is used to offer a low call dropping probability for multimedia connections. Another type of predictive routing scheme is proposed in [70] and [71], which exploits the predictive nature of LEO satellite topology to maximize the total number of users served by the system.

A satellite grouping and routing protocol is proposed in [69,73], which divides LEO satellites into groups according to the footprint area of the MEO satellites in each snapshot period. MEO satellites compute the minimum delay paths for their LEO members. There is a communication between LEO and MEO satellites, which consist of Delay report from LEO to MEO satellites, delay exchange in MEO, and routing table calculation. Another distributed routing algorithm is given in [69] which proposed a datagram routing algorithm for LEO satellites, which uses a decision map to define the allocated path. Another routing algorithm for a multicast satellite network is proposed in [73]. The figure 2.4.1 shows a grouping of satellites (group1 to group 3) and the traffic route from mobile terminal 1 (MT1) to mobile terminal 2 (MT2), which uses ISLs.

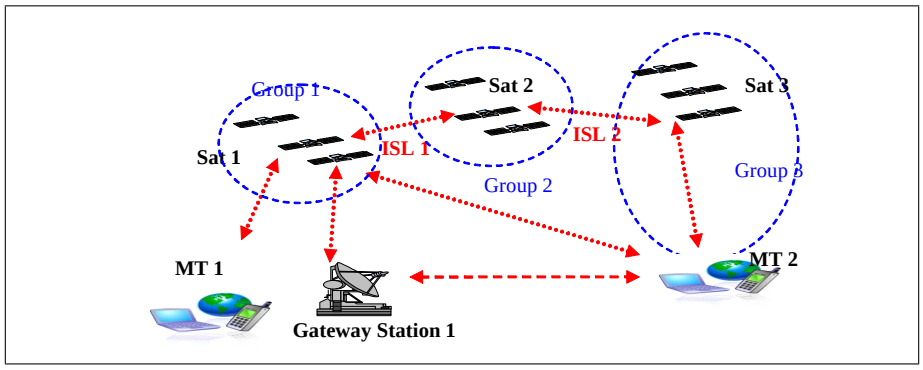

Figure 2.4.1: Satellite grouping and ISLs

2.5 Summary

In this chapter, some work that has been done in the field related to the topic of this thesis is briefly described. The three related topics are: traffic models, satellite networks and traffic allocation, which forms the basic for our research. Firstly research in the teletraffic field is reviewed, in which various types of current traffic models are studied. The current internet traffic falls mostly into the self-similar traffic model. However, since in this thesis, we focus more on the routing algorithm, we reduce the complexity of our research by using two types of traffic model: Poisson and Markov Modulated Poisson Process. Secondly, research in the field of satellite networks is reviewed, which gives us a global view of the satellite networks including their infrastructures and their signaling properties. Finally, the literature in traffic allocation schemes is reviewed. It varies from traffic allocations in terrestrial networks to our main topic research: traffic allocations in satellite networks.

Some important findings, which are related to our topic of research, even though some of these findings are not based on LEO satellite networks, were given: QoS in various types of traffic, dynamic properties of LEO satellites, inserting a clue in the addressing field, adaptivity, predictions, and distribution of traffic. These findings constitute important additional knowledge useful in our research, and will be discussed in further detail later in this thesis.

Chapter 3

TERRESTRIAL COMMUNICATION

Satellite communication will not take the place of existing terrestrial communication infrastructure. Instead satellite communication will fill in the gaps in the existing terrestrial network services. Therefore, an understanding of the terrestrial network is necessary. In addition, some technology in the terrestrial network might be used in the satellite network by introducing some modifications. First, we describe both wired and wireless technology, and then we focus more on wireless environments with their applications, including multimedia applications. Thereafter, we discuss the current Internet Protocol (IP), IPv4 and the future IPv6 since most research for routing in satellite network assumes IP as the protocol in a satellite network [68, 69, 72, 74, 75]. At the end of the chapter, an overview of satellite communication systems is given, which includes the topology and a taxonomy of such systems.

3.1 Background

3.1.1 Wired Technology

There are different options in building a communication infrastructure. The first option is to build it over the land - but wired over land installation projects often experience clashes with environmentalists, causing a delay in the installation. Another way to build a wired

communication is by installing it under the ocean. However, this method might have to go through a long and hard approval process, before operators can start to plant their cables under the ocean. This hindrance to the process will downgrade their competitiveness while their customers wait for these cables. These two cases are perhaps typical of the difficulties of building a wired communication infrastructure. In addition, wireless networks have their own characteristics, which make wireless networks preferable to wired ones. The most important advantages of the wireless network are mobility of the users and the relative lower cost of installations. In figure 3.1.1 an integrated network of wired and wireless technology is given.

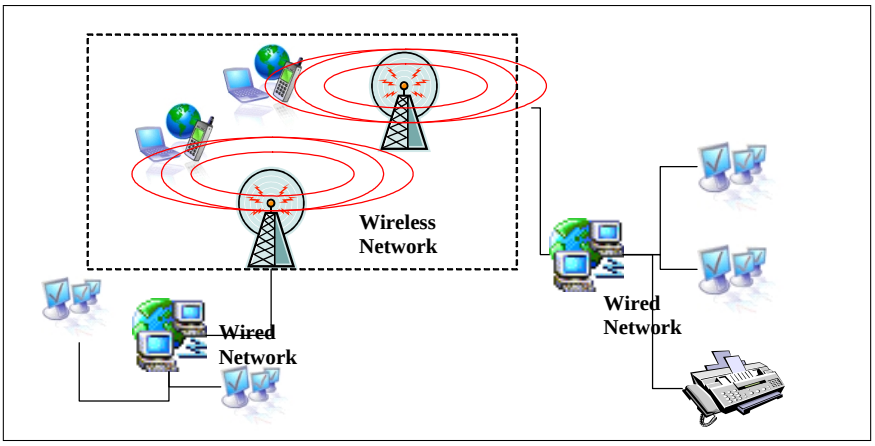

Figure 3.1.1: Interconnection between wired and wireless network

3.1.2 Wireless Technology

Because of difficulties in establishing a new wired connection, and due to the increasing need for broadband services and mobile systems, researchers try to find a solution by building a wireless communication. This approach, based on wireless technology, is expected to dominate the communications network market.

Since wireless technology will be used as basic knowledge for signal processing in satellite communication, we look at the history and basic ideas of this technology in the next paragraph.

3.2 Wireless Environment

3.2.1 History

The 20^{th} century witnessed the beginning of a wireless revolution. Globalization of the globe makes competition in every aspect of business greater. It follows that companies look to wireless technology for a strategic advantage. Furthermore, companies are trying to find a new market in rural areas. Although prices of wireless equipment and services have dropped, the market is expected to grow in size and scope of its services. Significant growth factors are:

1. Wireless technology is established and there is a lot of research in this area to support the development of this technology.

2. More companies would like to adopt wireless technology, due to the already established infrastructure.

3. Increased number of mobile users need communication wherever they are, even in rural or remote areas.

4. Special consideration is the rapid increase in the number of young generation users, who have become the largest wireless application users.

5. The number of multimedia applications and interactive applications is increasing.

All of these growth factors stimulate the interest of researchers and companies to explore further the prospects of wireless communication. There is a huge amount of research in wireless communication to find a better way of delivering data over wireless networks. The increased number of mobile users demanding multimedia communications is also increasing. A combination of mobile and multimedia is possible and affordable if a broadband service is provided by telecommunication system.With the increase in the number of mobile users, it is predicted that wireless carriers will be able to reduce the prices for both basic voice services and data communications. Consulting firm Ernst & Young reported that by 2008 wireless will overtake wireline as the dominant method of telecommunications worldwide [79].

Because of the development of broadband wireless communication, more communication companies are prepared to cope with delivering integrated voice and data communication ser-

vices to users and seek extra revenue from additional data services. They are trying to define the future model of wireless mobile communication. The development of mobile communication is given in chronological order by Cortese as follows [80].

The first generation of mobile communication was analog cellular, which carried only voice. In 1971 technical proposals for a cellular standard by AT&T and Bell Laboratory were written. This standard is called the Advanced Mobile Phone System (AMPS). It took about ten years before the first commercial cellular systems were introduced in the 1980s (using UHF 890 MHz in USA and 900 MHz in UK called TACS/Total Access Communication System which has a bandwidth of 75MHz for mobile telephony, and in Scandinavia Nordic Mobile Telephone/NMT which used 450 MHz and 900MHz). Other standards, as stated by Korhonen, were developed and used only in one country, such as C-Netz in West Germany and Radiocomm2000 in France [81]. The first commercial launch of the first generation mobile communication system was in 1983. Because of the lack of capacity at that time, the first generation mobile telecommunication system underwent a transition into D-AMPS (Digital AMPS) and then to the second generation of mobile communications.

The second generation of mobile communication is digital cellular. This can carry voice and data by introducing TDMA (Time Division Multiple Access), CDMA (Code Division Multiple Access), and GSM (originally Groupe Speciale Mobile, specifying the first standard 900 MHz Band with reserve block of 25 MHz in 1980 - Later changed to Global System for Mobile communication). Bekkers and Smits noted that digital transmission paths have several advantages compared to analogue systems. They improve spectrum efficiency, provide higher transmission quality and engage extensive security facilities (authorization, data encryption) [30]. According to Korhonen, the first launched of GSM network was in 1992. The second standard is the UK standard (1990). It uses two blocks of 75 MHz in the 1800MGz frequency band (Digital Communication Systems, DCS-1800), which is called PCN (Personal Communications Networks). In the USA, wireless network operators use a different frequency (1900 MHz) and it is called PCS-1900. Compared to the first generation of mobile telephony (AMPS), the second generation needs more sites due to the higher frequency. During the mid -1990s, several Asian countries started to apply digital standards such as: GSM, D-AMPS, and CDMA. In 1993, Japan announced a new frequency band standard for digital cellular systems (800 MHz and

1.5 GHz), which is called Personal Digital Cellular (PDC). Numbers of mobile subscribers with their technology in October 2002 is given below (Source EMC World Cellular Database [81]). According to this figure (figure 3.2.1), GSM based mobile communications is the strongest of the 2G models.

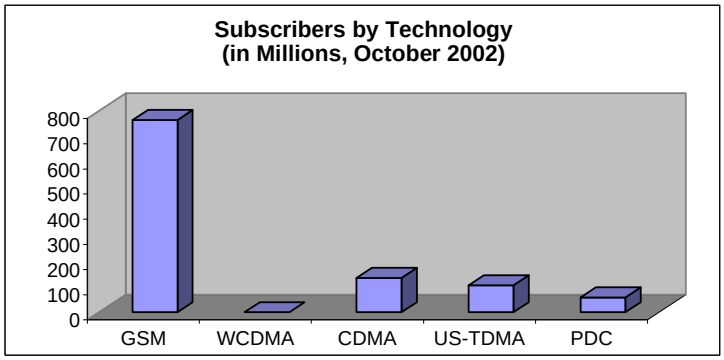

Figure 3.2.1: Total subscribers of different type of cellular technology in 2002

Generation 2.5 includes all advanced upgrades for the 2G networks. In this generation GSM has evolved with various improvements towards new systems: High Speed Circuit Switched Data (HSCSD), General Packet Radio Service (GPRS) and Enhanced Data rate for GSM Evolution (EDGE). The main reason is that the basic GSM could originally provide a 9.6 Kbps and later 14.4 Kbps data rate. According to Korhonen, HSCSD allows a mobile station to use more than one time slot for a data connection. The drawback is that because HSCSD is circuit switched, it allocates the used time slot completely. In GPRS, the data rates can reach 115Kbps maximum. Maximum data rate is available if an average time slot has 10Kbps and maximum of eight time slots can be used. EDGE uses a new modulation scheme called eight phase shift keying, which increases the data rates to three times that of GSM rates. EDGE can only be used over a short distance. A combination of GPRS and EDGE is called enhanced GPRS with maximum data rate of 384 Kbps [81].

The third generation is currently in the process of penetrating the wireless market. The transition from the second generation occurred because of concerns about the addition of new facilities and functions in the field e.g. high speed data transport, video, and multimedia. The third generation is to become the mobile extension of the fixed telecommunications in-

frastructure. One key requirement is higher bandwidth to allow high-speed wireless access to the Internet.

In 1997 both the Association of Radio Industries and Business (ARIB) and ETSI selected WCDMA (with bandwidth of 5 MHz) as their 3G radio interface candidate. Later, the most important companies in telecommunications joined forces in the 3GPP program, with the goal of producing a 3G system based on the ETSI Universal Terrestrial Radio Access (UTRA) radio interface. Since the radio spectrum allocated for UMTS is already taken in the US (IMT-2000 spectrum allocated to 2G PCS networks), in the USA operators are more attracted to CDMA 2000. Within Europe, Universal Mobile Telecommunication System (UMTS) will provide an initial transmission rate of 2 Mbps, over a transmission frequency of 2 GHz. This is despite the fact that UMTS introduced more technologies for the 3G wireless systems, which come under IMT-2000 (International Mobile Telecommunications) standards. The technical design for 3G standard is still under development. There are choices between 5 technologies: UTRA FDD (Frequency Division Duplex, which used Direct-Sequence spread spectrum CDMA), UTRA TDD (Time Division Duplex, also named as Time-Division CDMA) , hybrid FDMA/TDMA, CDMA2000 and a synchronous TD-CDMA called TD-SCDMA [7]. Operators in the USA planned to provide a service which will be available up to 10 Mbps for local area mobility services by 2005, but only in a provisional way. The final extension, everywhere, will be available by 2010. The development of mobile communication is given in figure 3.2.2.

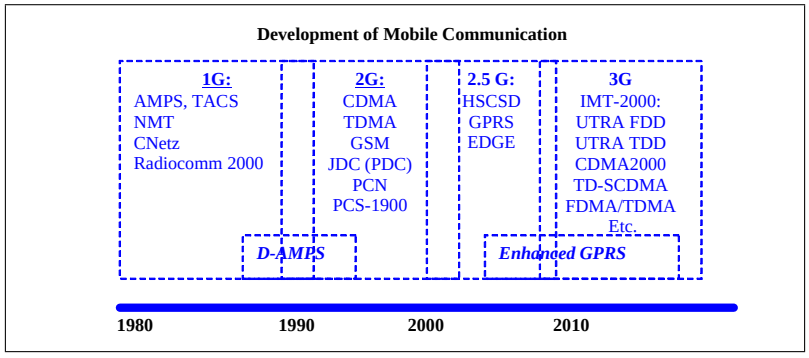

Figure 3.2.2: Development of cellular technology in mobile communication

Research into fourth generation technology has begun, which will probably be called software-defined radio or 4G mobile network. This generation probably will cover all modern technologies and enhanced telecommunication technologies. This 4G mobile network will integrate various types of communication technology such as wired technology, wireless technology and broadband satellite systems to form a global communication network as given in figure 3.2.3.

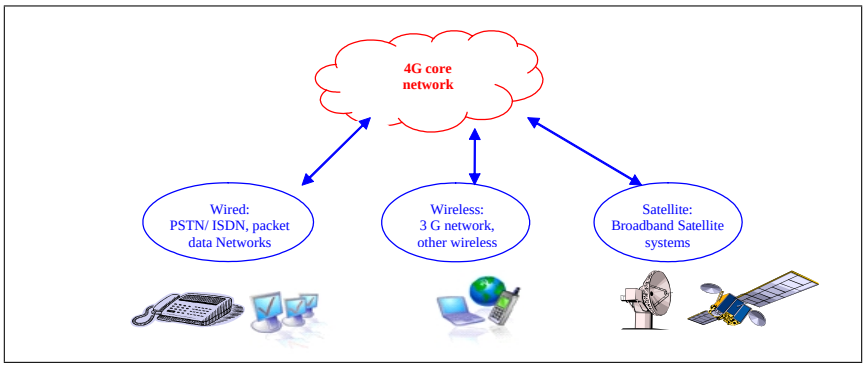

Figure 3.2.3: Hybrid communication systems

Currently, the wireless industry has focused in a large part on digital transmission technologies, which can improve spectral efficiency. Mobile radio systems work on the basis of cells for two reasons as mentioned by Webb in his paper [82]: The radio signal's strength decreases quadratically as the distance between the transmitter and the mobile user. Secondly, the radio-spectrum is limited. However, a new development in the design of cell based systems can help to improve the capacity of the network. Modern cellular networks are designed around the frequency reuse concept, where the system's area of operation is divided into many sub regions or cells, called clusters. The service provider can effectively multiply its available bandwidth many times by linking many clusters together within a network's area. If more bandwidth is needed, cells and cell clusters can be reduced in size in order to recycle the same frequencies over less area. The minimum size of the cell cluster is defined by the properties of transmission signals and their noise. The reuse distance indicates the distance between base stations using the same frequencies. Neighbouring cells use different frequencies, although cells that are further apart can use the same frequencies, as they are not hindered by each other due to the relative low power of transmitting stations.

In figure 3.2.4 a hexagonal pattern of a wireless cellular system is given, with reuse factor 7 (seven). Stallings defined the reuse factor as the number of cells in a repetitious pattern (each cell in the pattern uses a unique band of frequencies) [83]. Implementing this pattern in a wireless cellular system will increase the possible frequency reuse.

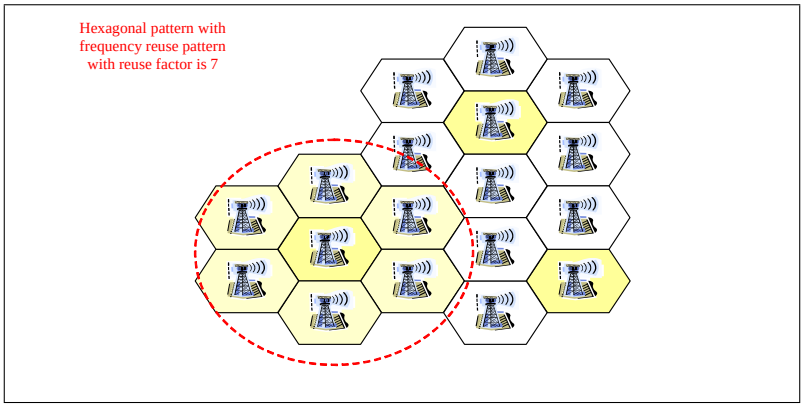

Figure 3.2.4: Frequency reuse in cellular technology with frequency reuse factor of seven

Urban environments are unfriendly to cell based wireless, due to signal obstruction, and multipath effects. This is the reason why the actual urban cell coverage zone is very poor compared to the theoretical coverage zone. Imprecise cell boundaries make it difficult to design an efficient clustering and reuse scheme. Thus, in urban regions, cellular systems operate with only a small fraction of their capacity. Hence, there is a need for a more sophisticated cellular design model, which will have cell coverage areas with effectively shaped. In addition there is a need to develop ways to have high levels of accuracy for the cellular channels.

There are difficulties in designing an effective cellular design model, because of the lack of accurate Radio Frequency environmental information. Only a few in-field average power measurements (surveys) have been done. Most of the system coverage surveys today are limited to computer simulation. This can be used for low density rural planning, but does not produce accurate data information needed to properly design an urban and semi-urban cellular model. Since significant information about the RF environment is not available, it is difficult to define the irregularity and actual cell size. Provided with only a few guidelines for the process of clustering zones, the cell site planners must define carefully the cell sizes and their overlap

coverage regions. If they do not design the cell size optimally, there is a possibility that it will result in the decrease in available bandwidth.

As the requirements for available bandwidth and required reliability have increased, new approaches should be used to evolve a new commercial standard for cellular design. Satellite remote sensing data need to be used to collect a more accurate field characterization of the wireless environment. The current cells are designed for an ideal geometry; however, they will be applied to a less than ideal environment. The resulting irregularities from a less ideal environment represent loss in coverage and user capacity. In the future, the cell's architecture must be designed for highly flexible cell geometry that can be implemented to the less ideal local environment. There is a barrier to optimizing the cellular architectures due to the site interference, which is usually site-specific interference. If cell sites are packed more densely with uncontrolled equipment and insensitive receivers, this will become a central limiting factor to the success of low tolerance next generation technologies. Therefore, design technologies must be complemented by improvements in filter efficiency and site management.

3.2.2 Multimedia Application

A multimedia application is an application which consists of several media components such as speech, video, still images, and music. Multimedia applications can be interactive or distributional. Due to recent developments in transmission and computing technologies, distributed multimedia applications now become possible, using wired and wireless networks. Demands for multimedia traffic will grow significantly in the future. These multimedia applications require a reliable environment in wireless and mobile network. Voice and video applications, which depend on timely delivery of data, suffer from high variance of bandwidth and bit error rate. Physical layer error included by path-loss, fading, channel interference, and shadowing become significant in wireless networks, compared to wired networks, which support constant transmission rates and typical, predictable values for BER. We need insight as to how a multimedia application is decoded, and to understand QoS requirements of this multimedia application.

Traditional video compression methods are designed for reliable wired networks. Korhonen [81] noted that a multimedia presentation (such as video) requires a large amount of data to be transmitted, which sometimes contains much redundant information. When the same

compression method is used in wireless environments, additional considerations arise. First, the channel coding has to provide reliability for the whole bandwidth of the video stream. However, this will be expensive. Also separating compression method (source coding) and channel coding could waste resources, because the channel coder cannot distinguish the importance levels of the stream data. Therefore, a combined source and channel coding could perform better than a separated approach. Another consideration is the frequent changes of channel conditions due to the multipath effect, signal obstructions, etc [81].

The transmission of multimedia applications in wireless environments requires a new video encoding. There is one proposed solution from Frankhauser et. al. Since the transmission of wireless video of acceptable quality is significant, they propose a video coding called Wavevideo. This is an integrated-adaptive video coding architecture, designed for wireless networks. The video coding includes basic video compression algorithms based on wavelet transformations, an efficient channel coding, a filter architecture for receiver-based media scaling, and error control methods to make this coding adaptable to the wireless environment [5].

In other research in multimedia communication, Wang et al. describe the performance of image communication via satellite network. In order to reduce the problems mentioned above, while designing a video transmission system for wireless networks the coding method should be robust. The physical and link layers of wireless network provide feature channel codes that can correct typical errors. However, the application itself must have the possibility for recovering errors from 'bursts', which cause loss of packets. Also, the Human Visual System (HVS) helps us reduce the complexity of designing the codes. Because HVS is more sensitive to features, like sharp edge or changes in brightness, than to gradients or colors, there should be a high redundancy in the video codes. This will help in recovering the transmitted video applications with errors. In addition, the code needs to consider the efficiency of the channel coding. It will be easy to improve transmission quality by adding extra channel codes, but it will increase the cost. In case of drastic changes in network transmission conditions (due to handover or congestion), a coder must be capable of adaptive scaling (resizing) the video stream. In the case of heterogeneous networks with different wireless cells, with different bandwidth, and different BER, an error control and QoS adaptation as close to the wireless link as possible is required [84]. In case of a multicast group transmission, it is important to deliver the requested

QoS to all subscribers. Each link of a distribution tree in a network should never transport a better quality than the maximum required in the following sub tree.

The International Telecommunication Union (ITU) has developed standards for multimedia applications. ITU-T Recommendation H.261 [85,86] addresses videophone and videoconference applications at bit rates of multiples of 64 kb/s. The H.261 coding is organized as a hierarchy of groupings. The video stream is composed of a sequence of images, or frames. According to Liou [86], each frame is organized as a set of Groups of Blocks (GOB). Each GOB holds a set of 3 lines of 11 macro blocks (MB). Each MB carries information on a group of 16x16 pixels (Luminance information is specified for 4 blocks of 8x8 pixels, chrominance information is given by two "red" and "blue" color difference components at a resolution of only 8x8 pixels) [86].

Another ITU-T recommendation is ITU-T Recommendation H.262. One code that is already following this recommendation is MPEG2. This recommendation specifies coded representation of video data and the decoding process required to reconstruct pictures. The basic coding algorithm is a hybrid of motion compensated prediction. Pictures can be either coded in interlaced or progressive forms [87]. Necessary algorithmic elements are integrated into a single syntax, and a limited number of subsets are defined in terms of profile (functionalities) and level (parameters) to facilitate practical use of this generic video coding standard. This recommendation is also called ISO/IEC 13818-2. ITU-T Recommendation H.263 addresses similar applications as in H.261 but at a lower bit rates (less than 64 kb/s) [88].

The above standards are for video applications only, which are part of ITU standards for multimedia conferencing communications, H.32x standards. H.32x defines which related standards should be used for: encoding of video and audio applications, multiplexing, control, multipoint, data, and communication interface. H.32 standard series is given by Zhao and Loh in their paper [89]. First standard is H.320. This standard is used for narrowband switched digital ISDN. The standard for Broadband ISDN, ATM, and LAN is H.321, while H.322 is a standard for guaranteed bandwidth packet switched networks. H.323 gives a standard for non-guaranteed bandwidth packet switched networks (Ethernet). Analogue phone system PSTN (Public Switched Telephone Network) uses H.324 standard, while H.324/C gives the standard for mobile communication. The last standard is H.310. This standard is used for broadband ISDN, ATM, and LAN (as H.321 but it uses MPEG2 for video and audio codering, and MPEG

for multiplexing).

In all of these H.32x standards the ITU-T T.120 architecture is used to address real time data conferencing. The recommendations specify the manner in which to distribute files and graphical information, efficiently and reliably in real-time, during a multipoint, multimedia meeting. The objective of T.120 standards is to assure communications between protocols and services in different layers.

3.2.3 Quality of Service (QoS)

Quality of Service is defined by the International Telecommunication Union as the collection of service performance which determines the level of satisfaction of users. The ITU-T Recommendation X.641 (ISO/IEC IS13236) provides the basic definition for the most important QoS concepts in the information technology environment. The main concept of QoS characteristics is to have a "quantifiable aspect of QoS which is defined independently of the means by which it is controlled or represented" [90].

Several international bodies have defined the frameworks for QoS as outlined by Espvik, Franken et al. The first organization is ITU-T (Recommendation E.800). This recommendation is adopted by IEC (International Electro technical Commission). The second organization is European Telecommunication Standards Institute (ETSI) framework, which is based on the work of the FITCE (Federation of Telecommunication Engineers of the European Community) Study Commission. International Standards Organization (ISO)/Open System Interconnection (OSI) and Telecommunication Information Networking Architecture Consortium (TINA-C) QoS framework contributes in layered and distributed architectures. In Europe, there is the European Institute for Research and Strategic Studies in Telecommunication (EURESCOM) [91].

Asensio and Villagra in [92, 93] differentiate QoS into different concepts: Aspects of QoS which are measurable are called QoS characteristics, while any information which is used in managing QoS are classified into requirement context and data context and are called QoS context. QoS requirements are information about all or part of a requirement to manage QoS characteristics; and are expressed in QoS parameters, e.g. a maximum value, a target, or a threshold. QoS relationship defines the mutual relationship between an object and its environment and includes the QoS requirements and the QoS offer which will be expressed in

the QoS contract. The set of QoS characteristics provided by an object is called QoS capability. The QoS offer describes the advertised QoS. The QoS contract defines the relationship between entities [92, 93].

The traditional network service on the Internet is a best-effort datagram transmission. Best effort means that there are no specific traffic parameters associated with the datagram and there is no absolute guarantee provided. For applications that tolerate packet delays and packet losses, this best-effort model will be satisfactory. ATM network uses also this best effort model. The Best Effort model handles two classes of traffic type in ATM network: Unspecified Bit Rate (UBR) and Available Bit Rate (ABR) type of traffic, which are typical for 'bursty' LAN traffic and data that is more tolerant of delays and cell loss. UBR is a best effort service that does not specify bit rate or traffic parameters and has no quality of service guarantees, while ABR is a managed service based on minimum cell rate (MCR) and with a low cell loss.

In view of increasing use of real time applications, which have very different characteristics and requirements than data applications, the best effort model might not be a suitable transmission model. Real time applications are less tolerant of delay variations and need specific network conditions in order to perform well. In ATM traffic model these applications which are real time based are classified into two types of traffic: Constant Bit Rate (CBR; traffic which is characterized by a continuous stream of bits at a steady rate, i.e. a low bandwidth traffic that is highly sensitive to delay and intolerant to cell loss) and Variable Bit Rate (VBR; there are two type of VBR: Real time/VBR-RT, wherein end-to-end delay is critical, such as interactive video conferencing and non real time/VBR-NRT wherein delay is not so critical such as video playback, training tapes and video mail messages). In order to provide support for these applications internet protocol is extended with the Quality of Service model.

Figure 3.2.5 gives the requirements of various applications in two variables according to Stallings [83]: Firstly, the sensitivity of these applications in time delays and secondly, the criticality of these applications. As given in figure 3.2.5 some applications are more sensitive to time delay (such as Personal voice over IP) and some are more critical in reliability (such as server backup). In addition there are some applications in which both of these requirements are significant (such as CEO videoconference and financial transactions) [83].

There are two architectures being defined to support this model. The first architecture

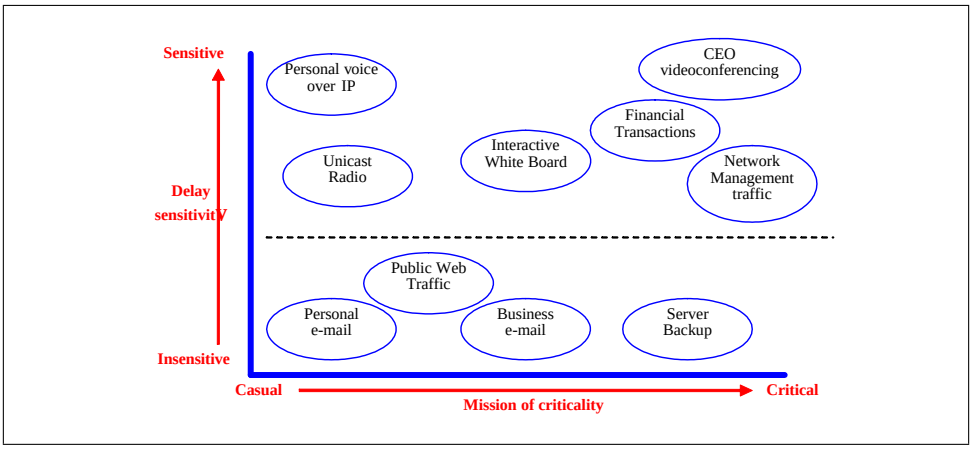

Figure 3.2.5: Different requirements for different applications [96] p.18

is called Integrated Services. This architecture provides applications with end-to-end QoS guarantees. The application specifies its QoS requirements and these are guaranteed by the network. The major drawback of Integrated Services is that the quantity of state information for each node is proportional to the number of application flows. Each traffic flow in this service can be classified under one of the three service classes, as given by Jamalipour [7]:

1. Guaranteed service class: provides for delay bound service agreements, which require critical delay constraints.

2. Controlled load service class: provides for a form of statistical delay service agreement.

3. Best effort service: matches the current IP service mainly for interactive burst traffic (such as web), interactive bulk traffic (such as ftp), and background / asynchronous traffic (such as e-mail).

One protocol that supports this integrated services architecture is Resource Reservation Protocol (also known as RSVP protocol), which provides the resource requirements from an application to the routers located between a source and destination hosts.

The second architecture is called Differentiated Services. In this architecture, traffic is classified into a finite number of priority or delay classes. This means that there should be no end-to-end guarantees; instead a privilege for a higher priority traffic class is given.

Information about priorities of traffic is given under Type of Service in IPv4 packet headers, or under the Traffic Class field in IPv6 packet headers. In general, there are some QoS parameters to be considered, such as timeliness (delay, response time), bandwidth, reliability [95].

3.2.4 Internet Protocol over Wireless Links

According to specifications of the Quality of Service, end users define the performance level they require. Performance of service depends predominantly on wireless system characteristics and transmission techniques that are used for transmitting data.

With increased use of Internet, the Internet protocol (IP) becomes more significant. Currently, a new generation of Internet protocol (IPv6) that is QoS sensitive has been used in several countries for Internet communications. Before we discuss the specific details of this QoS sensitive protocol, some additional information can help us to understand this popular Internet protocol. The offered QoS in a certain network depends on the protocol chosen in that network. According to Stallings [94], a protocol is "a set of rules or conventions that allow corresponding layers in two systems to communicate by means of formatted blocks of data, which obey this set of rules or conventions". The key features of a protocol are syntax (concerns the format of data blocks), semantics (includes control information for coordination and error handling), and timing (includes speed matching and sequencing) [94] p.72. Figure 3.2.6 shows different layers in TCP/IP protocols (internet stack), which match the standard of International Standards Organization (ISO) Open Systems Interconnect (OSI) reference model:

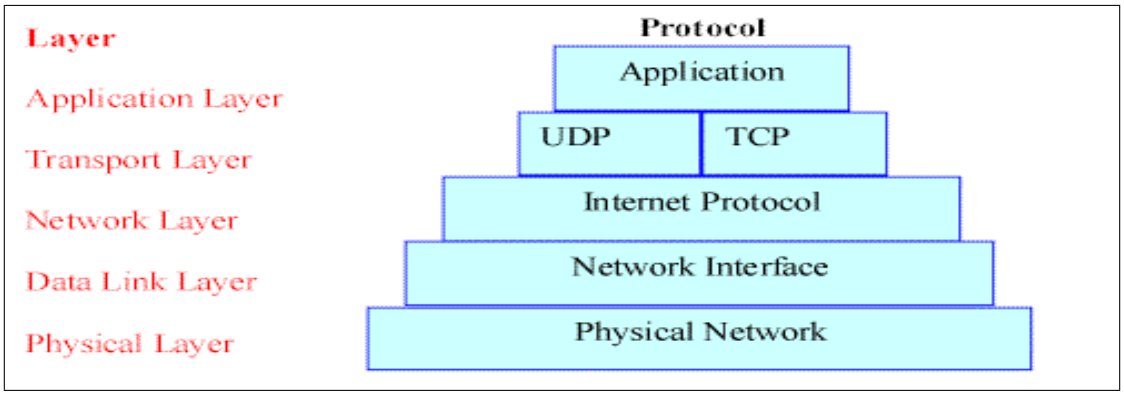

Figure 3.2.6: TCP/IP protocol architecture

In the transport layer, there are two types of transport protocols, TCP (Transmission Control Protocol) and UDP (User Datagram Protocol). According to Postel [96], TCP and UDP use the same addressing scheme, an IP address (32 bits number) and a port number (a 16 bit number). The IP address is used to route the packet to their host destination on a network, while the port number is used to route the packet to the right process on the host. UDP is described in STD-6/RFC-768. This protocol provides a connectionless host-to-host communication path, which has minimal overhead. Since it is connectionless, the datagram can be sent at any time without any negotiation or reservation. Therefore, there is no guarantee that the datagram will be received in the destination's host. This protocol is unreliable, but fast and can be used to broadcast data. Since it is unreliable, the application on the receiver host needs to be able to recover the missing datagrams [96]. TCP is described in STD-7/RFC-793 as a connection oriented protocol, which will provide reliable communication between two end processes. The unit data that is transferred is called a stream, which is constructed from a sequence of bytes [97].

If users send data from any application, TCP/IP receive it and divide it into small packets. A header will be attached into these packets to give information about the destination address. A packet is transmitted to the Internet network layer, which will enclose this packet into an IP datagram and add the datagram header. The internet network layer decides which route this datagram will follow before sending it to the Network Interface Layer. From this layer, datagrams will be sent as frames over the physical layer (network Ethernet or token ring networks). When the frames reach their host destination, the reverse procedure will follow until the user in the destination's host can use this data in their application layer.

In a wireless environment, there are several types of protocol in which the user can have internet access other than by a telephony service. In the previous sub-chapter we described the history of cellular wireless networks. Before 2.5G version, WAP (Wireless Application Protocol) as shown in figure 3.2.7 provided users of this cellular wireless terminals access to telephony and information services, including the internet and the web.

In addition to cellular wireless networks there are other wireless networks, which are used as either LAN or WAN. One of these is a Wireless Local Loop (WLL, with its IEEE 802.16 protocols) 3.2.8, which is sometimes called fixed wireless access (fixed subscribers). This WLL

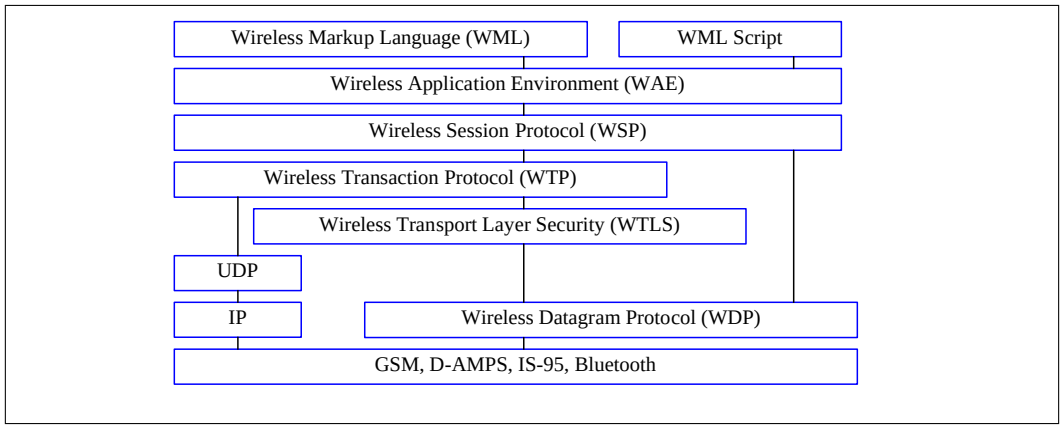

Figure 3.2.7: Wireless Application Protocol [83] p.401

provides an alternative solution for a wired network. Its advantages are e.g. reducing of installation cost and time, and it can provide a selective installation. Figure 3.2.8 illustrates this WLL protocol.

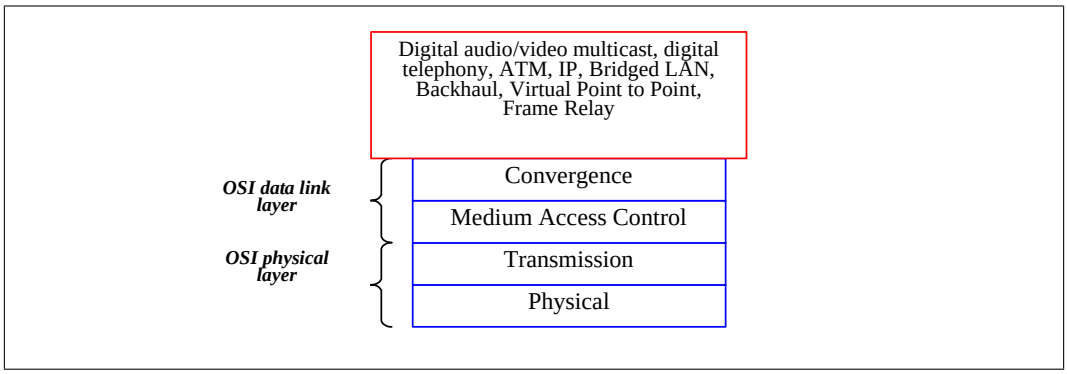

Figure 3.2.8: Wireless Local Loop (IEEE 802.16) [83] p.370

Another type of wireless network is Wireless Local Area Network (W-LAN), with its IEEE 802.11 protocols. In this W-LAN the subscribers are mobile. In figure 3.2.9 the protocol for this WLAN architecture is provided.

Due to the increased numbers of Internet users, current Internet Protocol, IPv4 (IP version 4) cannot cope with the growth in IP addresses. Another issue is that Ipv4 cannot accommodate the requirements from current applications, especially with limited address space and routing

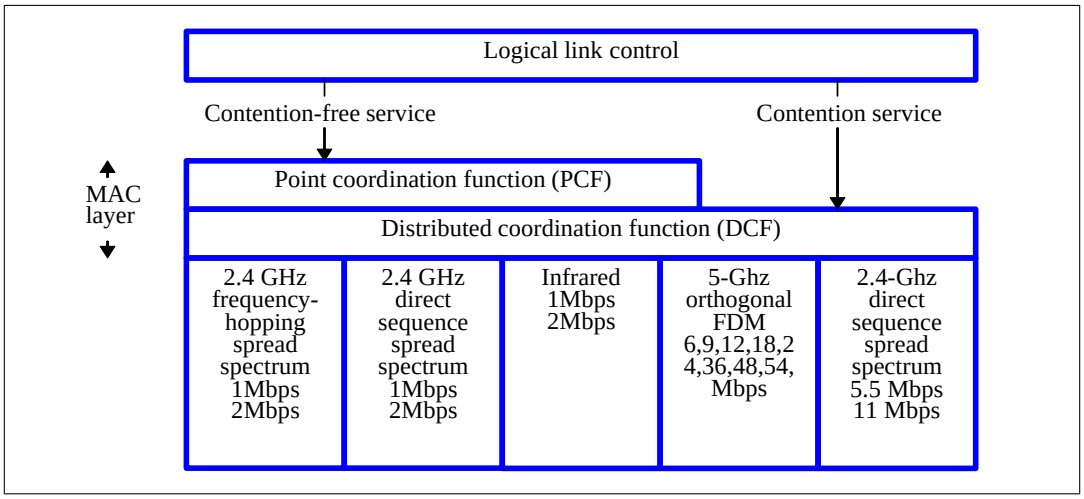

Figure 3.2.9: Wireless Local Area Network (IEEE802.11) [83] p.463

complexity. Therefore, IPv6 (IP version 6) starts replacing IPv4. The IPv6 provides 128 bit addressing instead of 32 bit addressing in IPv4 [98–103]. IPv6 has three types of addresses:

- Unicast: packet is delivered to the interface identified by this address.
- Multicast: packet is delivered to all interfaces identified by this address.
- Anycast: packet is sent by a single sender to multiple destinations, but only one destination will reply to the sender, e.g.: request of information from the nearest server, by sending a request to all servers.

The header format in IPv6 is simpler since the header length is constant, and there are no flags and no fragment offset. QoS in IPv6 improves by providing priority definition for packets in IP header (Priority field) [104]. There are two types of priorities [105]:

Congestion-controlled traffic: this traffic type responds to congestion by doing some alternative-limiting algorithm. The priorities are:

- 0: uncharacterized traffic.
- 1: non-delay sensitive traffic (e.g. background "filler" traffic/news).
- 2: unattended data transfer (e.g. email, SMTP).

- 3:reserved.
- 4: attended bulk transfer (e.g. FTP, NFS).
- 5:reserved.
- 6: delay sensitive traffic (interactive telnet).
- 7: network control messages (routing protocols, SNMP).

Non-congestion controlled traffic: this type of traffic drops packets when congestion occurs.

Currently, IPv6 has been implemented in some countries and will slowly replace the conventional IPv4.

In general, in a wireless environment, connectivity between cellular links and the Internet is as described by Xylomenos and Polyzos as given in figure 3.2.10. Inter Working Function (IWF) provides an inter connection between cellular telephony with other networks, while Radio Link Protocol (RLP) is used to enhance link functionality and performance between mobile users and the IWF [106].

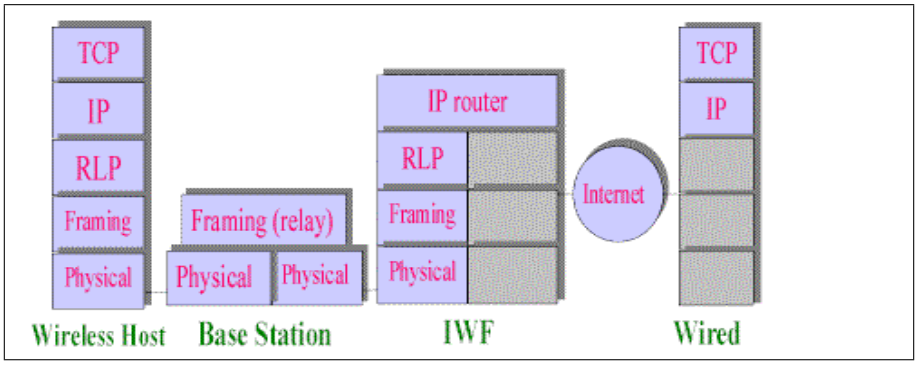

Figure 3.2.10: Internet connections in wireless environment [106]

3.2.5 Long Distance Communications and Communications in Rural Area

The wireless environment depends on wireless technology but also on the local cell-site environment. Requirements in urban areas are different from requirements in rural areas. Therefore, a specific study has to be accomplished before a flexible cell-site planner can be developed. On

the other hand, in some large countries, which have many rural and remote areas, communications become an important national policy issue. Borland in [107] mentioned that politicians are concerned about the growing gap between country and city, if telecommunication companies and cable TV companies are not investing in supplying high-speed Web access to remote and rural areas. Since remote areas will not deliver huge benefits for the telecommunication companies from providing such services to remote areas, they are not interested in building their infrastructure in these areas. By using wireless technology with a suitable infrastructure design, it is possible to support a policy of distributing communication infrastructure, equitably. This will provide an easy method of accessing information all over the world. This policy can be supported because the telecommunication company has fewer problems in setting up the infrastructure compared with wired technology, particularly when satellite communication technology can be brought into play [107].

Another problem is establishing a long distance communications between continents. It is very expensive to build communication towers or even relay stations with antenna towers in oceans. The earth's atmosphere can be used to reflect shortwave signals, but there is a limitation in bandwidth available.

Wallack in [108] mentioned that the problem of providing a high-speed network is not simply limited to the difficulties brought about by poorly served rural and underdeveloped areas. Sometimes, in urban areas other issues, such as environmental and social issues, make it difficult for the telecommunication companies to get a modem service installed quickly and easily. Therefore, another solution is available namely, using satellite network as the communication medium. Once the satellite infrastructure has been launched, installation time is often smaller than for other networks. A requirement is the installation of a wireless transceiver and antenna at the customer's site [108].

A satellite network will be able to support applications, such as international video conferencing, information and news gathering in rural areas that are not covered by terrestrial network. Satellites will be able to support point to point or multipoint links to transport data.

3.3 Satellite Communication

At the beginning of the satellite era, interest in satellite communication systems was in GEO satellites, because of the simplicity of controlling these satellites. Because of their heights, their orbital period is exactly one earth day. Hence, if the satellite is in an equatorial orbit and going in the right direction, it will appear stationary above the earth. This makes it very easy for a single satellite to serve a single geographic area.

A geosynchronous orbit is a good orbital position for spacecraft, as many benefits can be achieved. In this orbit, atmospheric drag and radiation effects are relatively small. Since GEO has already been used for many years, design technology, launching, positioning and switching for satellites in this orbit are well understood. Once the satellite is in its position, it is easy to control the position, and use it in a network. Because of this unchanging position, this satellite system can be seen as a fixed network topology. Due to this relative fixed position of GEO satellites in their orbit, there should be no control overhead to track the satellites. Network complexity is minimal, since there is no need to switch signals between satellites, and ground-station antennas need aim at only one point in the sky.

The concept of a geostationary orbit is not new. According to Graham, the concept of a GEO orbit was written about by the Russian theorist Konstantin Tsiolkovsky in 1903 [109]. In 1923, Herman Oberth, a German space scientist, wrote about sending rockets into interplanetary space and about space stations, which maintained their positions over the earth. Later, in 1928 Herman Potocnik, an Austrian army officer, designed a wheel shaped space station. He wrote an orbit at an altitude of 35,900 km as an orbit with a period equal to the earth's rotational period. This makes the space stations appear at a fixed point on the earth's equator [110]. The first person who proposed the use of this orbit for communications was Arthur C. Clarke. He published an article in "Wireless World" in October, 1945 titled "Extra terrestrial relays: Can Rocket Stations Give Worldwide radio coverage?" [111]. He accelerated world technology from the German rocket research of the day, to today's global communications via a network of three geostationary satellites around the earth's equator. He wrote in his article about the power needed for communications to and from space. He also calculated the impact of solar eclipses. According to Hogle, only in 1964, NASA launched the first geostationary satellite and satisfied the predictions of Clarke. The region of space used for placing geo-stationary satellites

is known as the Clarke Orbit and is situated some 35,767 km (approximately 22,275 miles) above the equator [112]. The Clarke belt is where the gravitational force of the earth is exactly countered by the satellites' centrifugal force.

Satellite networks have been employed to support various services. In communication networks, satellites function as relays in the sky for the terrestrial network. Satellites can function moreover, as a reference point in the sky, as navigation systems to help calculation a user's position. Also, satellites can be used to monitor the weather and earth sensing, in which the satellite will function as equipment to monitor the earth's surface and atmospheric changes. Satellites can record movement of equipment or physical objects on the earth surface. This technology is used in many geology applications and also for military systems. The most popular use of satellites is to support video broadcasting systems. Satellites broadcast television programs.

A Satellite system configuration can be divided into three segments, space segment, ground segment and user segment as shown in figure 3.3.1.The space segment is, basically, the satellite

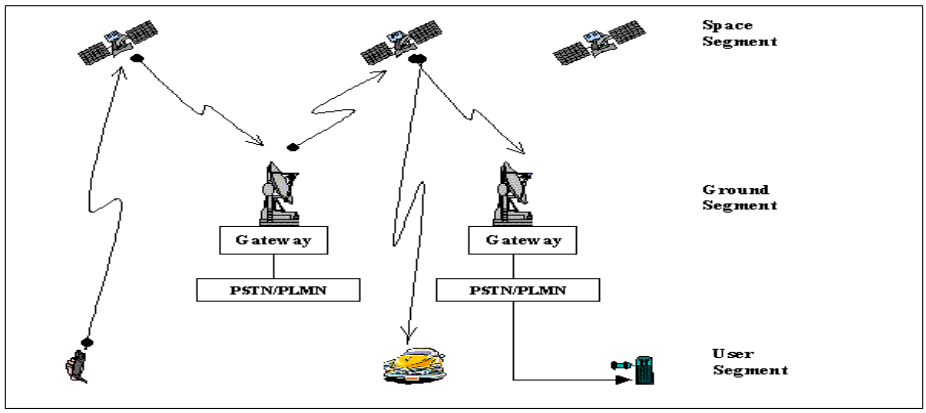

Figure 3.3.1: Segments in satellite communication

vehicle itself. In order to meet users' expectations a satellite based network must not only offer complete regional and global coverage, but it must approach the QoS that users expect from an existing terrestrial network. These requirements have implications for the architects of a satellite communications system. They influence a long list of choices that begin with the choice of where satellites will be placed in orbit. According to Carlson in his book [113], effectiveness

of a satellite application depends on the signal field strength of the link and the transmission path loss, which depends on the altitude of a satellite, its antenna system and effective radiated power (ERP). Basic transmission path loss, L_p is given by the following formula, wherein P_t and P_r are the transmitted power and the received power. G_t and G_r are the transmitting antenna gain and the receiving antenna gain.

$$L_p(in\,dB) = P_t(dB) - P_r(dB) + G_t(dB) + G_r(dB) \tag{3.3.1}$$

Transmission loss occurs because of the spreading of the transponder beam over a large area (multipath/scattering of the signal), and also, due to the fact that the signal is traveling through the earth's ionosphere and troposphere (separation between transmitter and receiver). Since the distance between a transmitter and a receiver is significant for transmission loss then, the choice of satellite orbit becomes important. The satellites' orbital geometry determines the satellite coverage, power constraints, the resulting dynamic network topology and round trip latency and variation. In circular orbits, satellites can provide a continuous coverage of an area inside their footprint. The footprint moves following the satellite movement. In elliptical orbits satellites only provide coverage when they move very slowly (in the apogee, the farthest position relative to the earth). While satellites are in the perigee position (closest to the earth's surface), they do not provide coverage, since the service is switched off. Another satellite, which is in apogee, will provide the coverage instead. The second segment is the ground segment. It consists of two elements. The first element is Network Operations and the Control Centre (NOCC) that provides network monitoring, configuration and control functions. Gateway stations are the second element, which function as repeater or switching stations. The Gateway station provides an interface with the public switched telephone network and communicates with mobile terminals. The last segment is the user Segment. These are the terminals or users, which can be in aircraft, cars, trains, ships, or personal handheld communication instruments.

The path in the GEO satellite system usually uses a simple bent-pipe architecture (up to the satellite and back down again). The satellite's function is to receive a signal from earth, shift it to a different frequency and/or different antenna beam and send it back to earth.

Current satellite communication systems are almost all based on geostationary satellites.

By using GEO satellite systems, a worldwide communication system can be designed. Some signal processing can be performed similar to signal processing in terrestrial communication networks.

In rural and underdeveloped urban areas, the market is suitable for satellite networks. According to Adamson, Smith et al. [114], this is especially the case for voice telephony - given the gaps in global coverage today as much as 50% of the world's population still has not made a phone call, due to lack of infrastructure. And, the International Telecommunication Union (ITU) states that 50 million people who can afford a phone can not get one (again, due to lack of infrastructure) [114].

Places like China, South America, Africa, Indonesia, parts of Russia and Australia inspire confidence that satellite communication will become the boom market predicted by investors. According to Ananasso and Priscoli, about 30% of all satellite communication will emphasize "cellular components", offering dual-mode functionality in which user instruments and ground systems make the decision as to which form of signaling-satellite or cellular-is necessary to complete a particular call [115].

The ability of satellite communications to complement terrestrial networks effectively is well recognized. This is particularly true wherever terrestrial networks are either not competitive (low traffic densities), not applicable (as in the case of maritime and aeronautical services) or less developed. Some of the features that will give satellite communications advantages over terrestrial communications are their wide area coverage of a country, region or continent, their independence from terrestrial infrastructure, their rapid installation of ground networks, their low cost per added site, their uniform service characteristics, their large channel capacity in each transponder channel and the benefits of the implementation of mobile or wireless communications in maritime and aeronautical applications.

FAA and COMSTAC forecast the future demand for Geo-synchronous orbit (GSO) satellites and Non Geo-synchronous orbit (NGSO) satellites. The actual number of satellites that has been launched between the year 1993 and 2002 is given and the demand for satellites from 2003 until 2012 is forecast, as given in figure 3.3.2 [116].

The number of satellites increased from 10 GSO satellites in 1993 to 28 GSO satellites in 1997. However, the number of NGSO satellites reached their highest number of 82 satellites

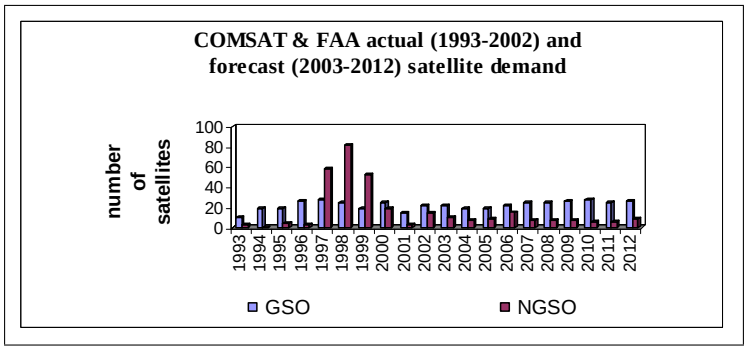

Figure 3.3.2: The actual and forecast satellite demand

in 1998. This was when Globalstar and IRIDIUM systems were launched. They forecast that the number of new NGSO system should not exceed this highest value in 1998, because of competing companies, slowing down in global economy, and especially due to the longer than the expected survival time of these NGSO satellites.

3.3.1 Various Satellite Network Systems

There is an international community of network designers, operators, vendors, and researchers who are concerned with the evolution of Internet architecture and the smooth operation of the Internet, which is called Internet Engineering Task Force (IETF). Inside this organization, there is a special project, which deals with the development of the mobile network. It is called the RACE MONET Project. They define some terms which are helpful for uniformity of understanding when discussing satellite communication.

Based on these definitions, it should be possible to integrate satellite communication systems with UMTS as given in the previous chapter. First, we need to select the most appropriate UMTS satellite system configuration, i.e. the most appropriate mapping of UMTS Network Entities (NEs) of the terrestrial communication system into satellite system physical entities: mobile terminal (MT), satellites, and Fixed Earth Stations (FESs).

According to Ananasso and Priscoli, following the type of switching procedures in the satellites, there are three different satellite configurations [115]:

1. Bent Pipe satellites (BP Sat) provided with transparent repeaters and no switching ca-

pabilities (e.g. Inmarsat-P, Globalstar and Skybridge). This type of satellite will perform as a space-based retransmitter of traffic received from user terminals and local Gateways in its footprint, returning the traffic to the ground. Communication between satellites and Gateway perform a bent-pipe channel. Satellite forms a wireless connection between nearby ground stations, which are Gateways to the terrestrial network.

2. Cross Connect Satellites (XC Sat) have on-board switching capabilities in the satellite. Some control is performed by some FESs. The second type of satellite will provide a network switch that is able to communicate with neighboring satellites by using radio or laser ISLs. IRIDIUM, Teledesic, Hughes Spaceway and Astrolink networks will use this types of satellite.

3. Intelligent Switching Satellites (SW Sat) with full on-board control (e.g. future satellite systems). In this class, there are on-board MSCP NE and possibly on-board MSDP.

In addition, satellite systems can be characterized based on three general categories of the services they provide according to Golding [23]:

1. Fixed satellite systems: This satellite system transmits to stationary earth stations that may require relatively large antennas based on specific traffic requirements, and are capable of supporting voice, data, and video applications including television distribution.

2. Broadcast satellite systems: These satellite systems use fixed earth stations for video or television distribution, which are typically smaller than those of fixed systems, but may use some MTs for audio applications.

3. Mobile Satellite systems: This satellite system supports a variety of low-bit rate services and typically uses MTs with small antennas.

Besides the above ways of describing satellite configurations, a satellite's configuration can be specified following their topology parameters: number of satellites in the satellite constellation, orbital attitude of these satellites, number of planes on which the satellites travel in their orbit, orbital period of the satellite, number of buffers in one satellite, and number of ISLs.

In this thesis, we focus on satellite systems which support communication networks. There are four different satellite communication groups depending on their altitude and their type of orbit. Figure 3.3.3 shows these four different types of satellite communication:

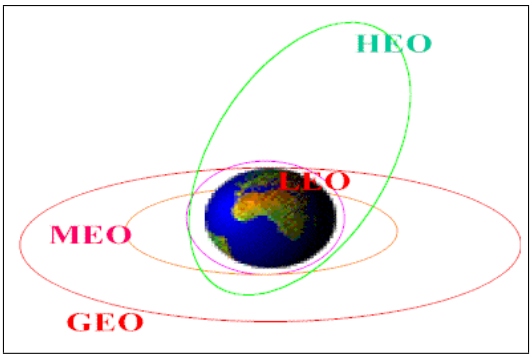

Figure 3.3.3: Four different orbital position of satellites

1. Geo-stationary Earth Orbit (GEO) satellites: large satellites in geosynchronous orbit should be used. GEO satellites have circular orbital altitude about 35,767 km above the earth's equator. Large antennas and higher carrier frequency improve the system's capacity. The coverage of a GEO satellite will only be possible when the latitude of the satellite is less than 75° according to Cruickshank, Sun et al. [52]. Consequently, to provide a global coverage, a minimum number of 3 satellites must be used. GEO satellites have been used for voice communication (e.g. Inmarsat -global coverage, AceS-single satellite for targeted region), and for broadband data (eg. Hughes Spaceway, Loral Astrolink). The propagation delay for a round-trip is about 540 ms [117].

2. Medium Earth Orbit (MEO) satellites: smaller satellites that have an orbit between 9,000km and 11,000 km above the earth, which reduces the transmission delay. The propagation delay between ground station and satellite is less than 80 ms for a round-trip. MEO provides services for voice (e.g. ICO satellite system), and were proposed for broadband data (eg. Orbilink and Hughes Spaceway NGSO) [52].

3. Low Earth Orbit (LEO) satellites: like MEO by using small satellites that orbit between 500 km to 1500 km above the earth. Because of their lower altitudes those satellites have

a shorter time period than the GEO satellites, and also, a lower delay (a round trip delay about 15 ms between ground and LEO satellites). LEO provides services for voice (e.g. IRIDIUM, Globalstar), messaging (e.g. Orbcomm) and proposed broadband data service (e.g. Teledesic, Skybridge, and the next generation of IRIDIUM-IRIDIUM Next/INX).

4. Highly Elliptical Orbits (HEO): these satellites have an elliptical orbit in contrast to the previous three circular orbit satellites. A satellite in this HEO system, generally, will only have coverage when they are near apogee (the furthest from earth's surface), while the satellite moves relatively slowly. When the satellite moves from near high apogee to low perigee (lowest from earth's surface) the speed is increasing and it has a small coverage. Generally, at this time the service will be disabled and another satellite in the constellation which is nearing the apogee will provide the service. To protect a satellite from the Van Allen radiation belts, the system is shut down. The orbits are generally at an inclination of 63.4° and provide carefully targeted, select rather than general, global coverage. HEO will provide services for broadband data (as proposed in Virtual GEO and Pentriad), which has apogees beyond the geostationary orbit, with a high delay. The virtual GEO plans to have an ISL between the satellites at apogee of different elliptical orbits.

One more orbital type of satellite is mentioned in FAA and COMSTAC report [116], which is called External (EXT) Orbit. In this orbit, the satellites follow a non-geocentric orbit and are centered on a celestial body other than earth (for example the moon). However, this type of orbit has not been considered as an orbit for communication systems purposes. Satellites in LEO and MEO can also be classified following the shape of their orbits into two groups according to Ananasso and Priscoli [115]:

1. Elliptical orbit: In this type of orbit, the earth is located in one of the two focal points of the elliptical orbit. Elliptical orbits have a longer visibility period of satellites over the highly populated areas, because the speed of the satellite is lowest when it is located farthest from the earth and highest when it is located closest to the earth. Examples of this elliptical orbit are Ellipso and Molniya.

2. Circular orbit: In this type of orbit, the earth is located at the center of the orbit.

Therefore, the altitude of the satellite from the earth's center is constant during satellite motion. The speed is constant during the rotation. Examples of this circular orbit are IRIDIUM and Teledesic.

Both orbits have an associated inclination angle. This inclination angle is an angle at which a satellite orbit is tilted relative to the earth's equatorial plane. If the inclination angle is 90^0, the orbit is called a polar orbit (IRIDIUM and Teledesic are polar circular orbits). Others orbits will be referred to as inclined orbits (Globalstar is an inclined circular orbit). Polar orbits intersect over the poles. In a satellite system with circular polar orbits, the network resources are inefficiently utilized. In Polar Regions, circular orbital satellite systems will provide maximal coverage. Network efficiency could be improved by using inclined circular orbits.

Other than categorizing satellite constellation according to their satellite's switching procedure, the orbital altitude, and the shape of their orbital, there are other categorizations possible. Satellite constellation can be categorized according to their frequency bands used for services ($C-, L-, Ka-$,or Ku-band); by their intended service provided (either for voice, telephony broadband data, navigation, or messaging); or their terminal type (fixed or MTs) [118]. Different frequency bands are given by Rossum as in the following table (3.3.1) [119].

Table 3.3.1: Various frequency bands

Frequency Bands	**Frequency**
L-Band	0.5GHz to 1.5GHz
S-Band	2.4GHz to 2.8GHz
C-Band	4GHz to 8GHz
X-Band	8GHz to 9GHz
K1-Band	10.95GHz to 11.75GHz
K2-Band	11.75GHz to 12.50GHz
K3-Band	12.50GHz to 12.75GHz
Ku-Band	13GHz to 17GHz
Ka-Band	18GHz to 31GHz

Moreover, we should distinguish between definitions of Geosynchronous and Geostationary orbit. Geosynchronous orbit (GSO) is any orbit, which has a period equal to the earth's

rotational period. The earth's rotational period is not the same as one mean solar day (24 hrs); but it is the time the earth needs to make one rotation in inertial space (or fixed space, because the earth moves relative to the sun). It is equivalent to 23 hours, 56 minutes, 4 seconds of mean solar time. All geostationary orbits (GEO) must be geosynchronous and must be circular and have a zero inclination (geosynchronous satellites can have an inclination such as 20^0, and will then move north and south during orbit, while geostationary satellites cannot). The advantage of the GEO satellite system is that it remains stationary relative to the earth's surface. This makes this orbit ideal for communications. However, it has some drawbacks: first the long distance between the satellite and the ground; secondly, the limitation of the geostationary orbits (not only space, but also, the limited frequencies allocated for the up and down link).

Because of these drawbacks of GEO, researchers tried to find another solution by using MEO and LEO orbits. Due to the existence of two Van Allen radiation belts (one belt is located between 1500 and 5000 km, while the second belt is located between 13000 and 20000 km as given in figure 3.3.4), which contain trapped electrons and protons above the earth's atmosphere, the orbital space is classified in those two orbits (MEO and LEO). MEO satellites are located between two radiation belts and LEO satellites are located below the lowest radiation belt [58].Satellite orbital motion follows Kepler's Laws and Newton's Laws of motion and gravitation.

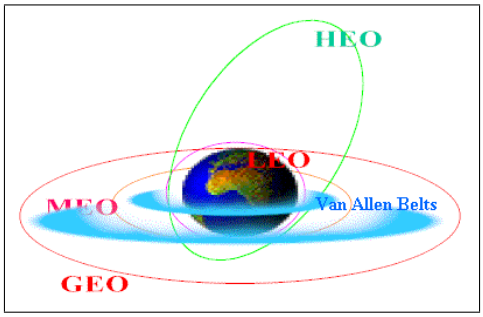

Figure 3.3.4: Satellite orbital and two Van Allen Belts

Kepler's laws described the motion of planets around the sun; these results are also valid for our satellites around the earth [120]. According to Newton's Law of Universal Gravitation, the gravitational force of attraction between two objects is directly proportional to the product of their masses, and inversely proportional to the square of the distance between them. Combined

with Newton's law of motion they form the physical basis for all theoretical work. The first law of motion of Newton states that every object will continue in its state of rest or in uniform motion in a straight line unless it is forced to change that state by forces impressed upon it. The second law states that the change of momentum measured relative to an inertial reference frame, is proportional to the force impressed upon it, and is in the same direction of that force. In our satellite system, we are more interested in the motion of a satellite around the earth. Once we identify the current state of a satellite's orbit (satellite's position and velocity) and the forces acting upon it, it should be possible to determine the satellite's state at some future (or past) time.

A satellite remains in its orbital position if the centripetal force holding a satellite in its circular path around the earth is equal to the gravitational force between the satellite and the earth. In circular orbit, the centripetal force (F_c) can be calculated using the equation:

$$F_c = \frac{4m_{satellite}\pi^2 R_{sat}}{T^2} \tag{3.3.2}$$

where R_{Sat} is the radius of the satellite orbit in meters, which is equal to the sum of the average equatorial radius of the earth, R_{earth}, and the altitude of the satellite, h_{sat}; T is the periodical time of satellite in second, and $m_{satellite}$ is the mass of satellite in kg. The gravitational force between the satellite and the earth (Newton's Law of gravitational Force) is

$$F_g = \frac{G(m_{satellite} \times m_{earth})}{R_{sat}^2} \tag{3.3.3}$$

where G is the universal gravitation constant ($6.67300 \times 10^{-11} m^3 kg^{-1} s^{-2}$), m_{earth} is the mass of earth in kg. In order to find the GEO orbit, the centripetal force will be equal to the gravitational force and the period is approximately 24 hours.

Consider the satellites and earth as our system. Our system can become very complex if we include all aspects, such as the fact that the earth is not uniform in density, and the fact that earth is not perfectly spherical. In order to reduce the system complexity, we include only two gravitational masses in our model, earth and the satellite itself, but we exclude the sun, moon and the other planets. The gravitational pulls of the sun and the moon have a negligible effect on LEO satellites, but not in GEO satellites. On the other hand, for some orbital classes of

satellite, there is an atmospheric drag, especially in LEO satellites [117]. The complexity of our orbital model depends on the level of accuracy and complexity required. Firstly, we need to determine how the accuracy of the prediction of the satellite position we require. In this case, we need to determine the types and relative magnitudes of forces that have significant effect in our orbital model (we only include e.g. the gravity of the earth and exclude corresponding non-uniform distribution of mass, gravitational attraction of the sun, moon, and planets, and atmospheric drag.). Furthermore, we need to define the complexity of the computational model we would like to have. There are two computational methods available. The first one is a numerical integration. This method starts with a satellite's position, velocity, and the sum (or integration) of all the forces acting on a satellite. This method provides accuracy but requires high calculation time to calculate the satellite's position and velocity for each time step (between known initial conditions and a desired prediction time). The second method provides an analytical solution i.e. a solution wherein, if we know the time of interest, we can directly calculate the state of the satellite's orbit at that time, without the need to integrate over time. It is used by North American Aerospace Defence Command (NORAD) and NASA's distributed model known as Simplified General Perturbation (SGP) as given by AMSAT. This reduces calculation time. SGP works by using data for all satellites on a daily basis (instead of all other models that only have orbital data for a limited number of satellites, such as the space shuttle). SGP has a special format of data, called mean Keplerian orbital element sets (two line orbital element sets) [121].

If other commercial satellite tracking packages require the same accurate predictions, they need to implement this format.A satellite network with accurate orbital configuration can provide better coverage for a specific region, or even more for global coverage. The proposed satellite constellation system will provide higher transmission rates and global coverage. In the future, some satellite constellations will offer up to 155Mbps (Teledesic).A satellite's position in its orbit is defined by using seven satellite orbital elements, which are called "Keplerian" elements. These numbers define an ellipse, the orientation of the satellite about the earth, and the place of the satellite on the ellipse, at a particular time.Basic orbital elements are given in AMSAT [121]; some elements, which are significant for our research, are:

1. Orbital Inclination: This inclination is an angle between the orbital plane of the satellite

and the equatorial plane. The orbital plane is a plane in which the orbit lies and always goes through the center of the earth, but may be tilted with any angle relative to the equator. The inclination is by convention a number between 0^0 and 180^0. Orbits with inclination near 90^0 are called polar (because the satellite crosses over the north and south poles). The intersection of the equatorial plane and the orbital plane is a line, which is called the line of nodes.

2. Right Ascension of Ascending Node (R.A.A.N): Using the inclination, we still have more than one orbital plane. Therefore, we specify the lines of nodes in the equator lines. The line of nodes occurs in two places: ascending node (where the satellite crosses the equator, when they move from south to north) and descending node (in the case where the satellite moves from north to south). An astronomical coordinate system called right ascension/declination coordinate system is used to define the RAAN. First, the definition of Vernal Equinox is a reference point in the sky which has a right ascension of zero. RAAN (Right Ascension of Ascending Node) is an angle, measured at the center of the earth, from the vernal equinox to the ascending node. By convention, RAAN is a number between 0^0 and 360^0.

These parameters are given in figure 3.3.5.

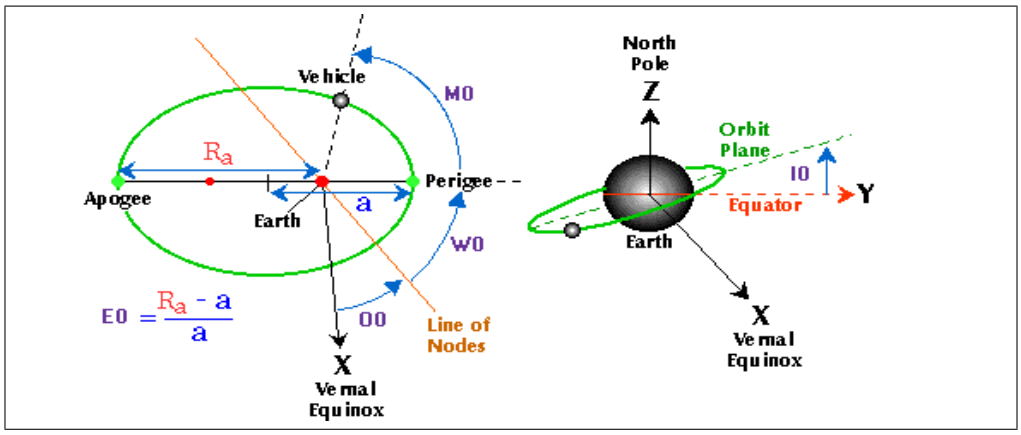

Figure 3.3.5: Keplerian Elements [122]

3.4 Summary

In this chapter, the advantages and disadvantages of wireless technology are studied. The wireless technology came into the telecommunication world by providing some benefits in contrast to the wired technology. Some of these benefits are lower installation cost, flexibility, and mobility. Similar features of terrestrial wireless technology arise in our LEO satellite network. Different types of wireless technology and their historical background are given which allow us to consider a more global and hybrid communication network. An integrated communication system of PSTN, WLL, WLAN, 3G (or 4G), and satellite communication (for a long distance or rural area) can be achieved by allowing each interface of these systems to reach an agreement in their 'protocol'. The current trend is the implementation of IP based satellite network for global communications. especially since the already started IPv6 can cope with the QoS requirement of the users, which is significant for the increases demand from the multimedia applications. The type of traffic class field in IPv6 and the new coding methods for multimedia applications provide us with the possibility of using this information to allocate multiservice traffic into the LEO satellite network.

The information about wireless technology in this Chapter and in combination with the information from Chapter 2 gives us the complete background to discuss satellite communication itself. In this Chapter a detailed description of a satellite network is given. Different classification of satellite's networks, depending on their satellite's switching procedures, their satellite's orbital altitudes and the shape of their orbitals are discussed. Two basic orbital elements from Keplerian elements will be used to define the position of a satellite in our simulation. In this thesis we consider only a LEO satellite constellation with onboard switching ability, which has a circular orbital.

Chapter 4

LEO SATELLITE COMMUNICATIONS

4.1 Introduction

In this chapter, LEO satellite communication will be discussed in detail. First the architecture of a LEO satellite constellation with its dynamic topology will be discussed. Issues in ISL and the handover mechanism will follow. Thereafter, some description of signaling issues, protocols and throughput parameters of the LEO satellite system will be given.

4.2 LEO Satellite Topology and Architecture

4.2.1 Differences between GEO and LEO

As mentioned in the previous chapter, satellite systems based on a geostationair satellite have some drawbacks. These drawbacks have led architects to consider two alternatives for the orbital position of the satellites: MEO and LEO. Their orbital position is lower than GEO satellites. LEO satellites appear to move constantly over the surface of the earth. MEO satellites typically have an orbital period of about six hours and LEO satellites have about 90 minutes. From the point of view of an earth-bound observer, MEO and LEO satellites are continually flying across the sky and disappearing. MEO and LEO satellite systems deliver some advantages when weighed against GEO satellites. Because their orbital positions are relatively

closer to the earth, the transmission delay will be less than that of GEO satellites and a lower power transmitter can be used (important for handheld components). The low altitude of LEO satellites that continuously fly by a geographic area may provide better overall coverage than a GEO. A LEO satellite system might pick up the call, even if there is an obstacle in the Line of Sight (LOS), because the satellites are numerous and offer a diverse path. If inter-satellite network capability is available, the number of Gateways required for global coverage can be reduced. This will cut down the tail charges and will enhance system reliability.

The most significant benefit of the LEO satellite system is that there is more reuse of limited available frequency, since the LEO satellite footprints are smaller than GEOs. Most of the commercial LEO satellite systems have an objective to reuse the limited allocated frequency, as much as possible. This reuse leads to a higher capacity of the LEO satellite systems. Frequency allocation of a satellite system is decided globally by the World Radio Congress (WRC) at its two yearly meeting; on the other hand the US Federal Communications Commissions hold the auctions of targeted frequency allocation in available bands.

Although these advantages make LEO satellites preferable to GEO satellites, the changing positions of LEO satellites present several problems in their implementation. Extra control overhead is required to track the movement of the satellites and to perform handover procedures.

Depending on their orbital altitude, LEO satellites may only be visible for a few minutes; therefore, many satellites are required to provide continuous communications. The higher velocity of LEO satellites compare to GEO satellites makes signal processing in LEO system more complex. However, these LEO satellites promise an important innovation for the mobile satellite system.

There are a number of major differences between the design of the GEO and the LEO satellite systems. The GEO system uses a static switching system and requires only a single hop in order to establish communication. The LEO satellite system in contrast uses a constant moving switching network and requires a multiple hop in order to establish a connection between a caller and a destination.

It is important to note that while a GEO satellite uses fuel mainly to maintain the precision of its station keeping position, a LEO satellite requires more fuel to maintain orbital altitude. Moreover GEO satellites are bigger than LEO satellites due to a large transmission power

needed. In order to launch a heavy GEO satellite into their orbital position, more launching procedures needs to be considered than launching a LEO satellite. The economic trade-off between weight (fuel needed, size of satellites, etc.) and cost (manufacturing cost, launching cost etc.), results in a design life of about 10 to 20 years for GEO satellites and 5 to 10 years for LEO satellites. Another difference is that the variable latency or jitter (variations in delay) that can cause problems in packet reordering at the destination is higher in LEO satellites than in GEO satellites. Because of the low orbit of LEO satellites, they may spend only a few minutes over a certain geographical area, which means a given transmission may be picked up and passed on by multiple satellites. Since satellite orbits are typically maintained within a range of locations, rather than one precise location, the pieces of a single transmission can be subjected to varied delays and subsequent packet reordering. The jitter can be readily cleaned up by creating larger memory buffers in earth stations and the transmission can be delayed long enough so that the playback to the user is at a constant latency.

4.2.2 Overview of LEO Satellite Constellation

A satellite constellation consists of a number of satellites in a large number of possible useful orbits, which deliver services to the users of this system. Preference is given to regular constellations, where all satellites share the same altitude and orbital inclination to the equator, in order to minimize the processing complexity. According to the types of services that LEO satellite system supports, the LEO satellite system itself can be categorized into two different groups; namely Little LEO and Big LEO.

Little LEO is named by Federal Communications Commission (FCC), because it uses a comparatively lower frequency than Big LEO [116]. Little LEO satellite systems use a spectrum lower than 1 GHZ to enable the use of lower cost transceivers. FCC has allocated frequency bands of 137-138MHz for downlinks and 148-149.9 MHz for uplinks to these systems [117]. Little LEO satellite system provides narrow band data communications (e-mail, two-way paging, and limited access to non-voice services). A wireless communication company, ORBCOMM, obtained its license for the LEO satellite system in 1994 and launched most of its satellites in 1997 and 1999. ORBCOMM will expand its constellation and plans to start launches in 2006. ORBCOMM constellation already has 48 satellites (35 satellites have been launched) at

825km and is already operational to cover a USA national service. Another Little LEO satellite constellation is under development such as given by FAA and COMSTAC report in table 4.2.1 as given in [116].

Table 4.2.1: Various Little LEO satellite constellation

SATELLITE SYSTEM	OPERATOR	SATELLITES			INITIAL LAUNCH
		NUMBER + spares	MASS (KG)	ORBIT TYPE	
ORBCOMM	ORBCOMM global LP	48	43	LEO	1997 (operational)
FAISat	FINAL ANALYSIS	26+6	151	LEO	1997 (under development)
LeoOne Worldwide	LEO-one USA	48	125	LEO	Under development
E-Sat	E-Sat, Inc ALCATEL	6	113	LEO	Under development

The second LEO satellites group is Big LEO satellite systems. Big LEO satellite system provides mobile voice telephony and data services in the 1-2GHz frequency range, namely 1610-1626.5 MHz for uplinks and 2483.5-2500MHz for downlinks. In table 4.2.2 various Big Leo satellite constellation, which are already operational and still under development are given in [116].

Another Big Leo satellite constellation, which is still under development, is Teledesic. In 1994, Teledesic designed an 840 broadband LEO satellite system, but in 1998 reduced the number of LEO satellites to a 288 satellite system, and more recently amended the system to 30 satellites with 3 spares in MEO orbital. ITU has set a deadline in September 2004 for Teledesic to operate the system.

Two Big LEOs have been fully deployed to date: IRIDIUM (66 satellites and 14 spares, at 780km) and Globalstar (48 satellites and 8 spares, at 1414km). Both of them provide a global service and variety of services, including voice, data, facsimile, paging and Radio Determination

Table 4.2.2: Various BIG LEO satellite constellation

SATELLITE SYSTEM	OPERATOR	SATELLITES			INITIAL LAUNCH
		NUMBER + spares	MASS (KG)	ORBIT TYPE	
GLOBALSTAR	Globalstar LP	48+6	447	LEO	1998 (operational)
IRIDIUM	IRIDIUM Satellite LLC	66+14	680	LEO	1997 (operational)
ECCO II	Constellation Communications	46	585	LEO	FCC License is given in July 2003 (Proposed)
Ellipso 2G	Mobile Com. Holding (MCHI)	26	1315	LEO& HEO	FCC License is given in July 2003 (Proposed)
Globalstar GS-2	Globalstar LP	64	830	LEO	FCC License is given in July 2003 (Proposed)
IRIDIUM / Macrocell	IRIDIUM Satellite LLC	96	1712	LEO	FCC License is given in July 2003 (Proposed)

Satellite Services (RDSS) to hand-held terminals. The differences between these two Big LEOs are listed below:

1. IRIDIUM uses Time Division Multiple Access (TDMA) and Time Division Multiplexing (TDM) which uses only one band for both uplink and downlink; while GLOBALSTAR Code Division Multiple Access (CDMA).

2. IRIDIUM has on-board processing and uses ISLs to perform a path from origin to destination; while GLOBALSTAR uses the bent-pipe approach to route long distance calls.

According to Wood, till now only IRIDIUM and Teledesic belong to satellite network constellations. Since only in these two satellite constellations the ISLs exists, wherein flexibility in routing the loaded traffic is independent of the terrestrial network. Onboard processing in the satellite allocates the traffic without a need to hop from one ground station to another [123].

In figure 4.2.1 illustrations of a bent-pipe based satellite constellation and a satellite constellation network with ISLs are given. A demand for a connection from mobile user 1 to mobile user 2 is required in both of cases. In the bent-pipe based satellite constellation, traffic from mobile user 1 is received by satellite 1 using the uplink channel. This traffic is then transmitted using the downlink channel to the Gateway Earth Station 1 (ES1). ES1 relays this traffic to ES2 and then to ES3 before the traffic is either sent directly to the mobile user 2 or sent via

satellite 3 to mobile user 2. Globalstar has this type of satellite constellation. In 2003, Globalstar received a FCC-license for a new LEO satellite constellation, which is called Globalstar GS-2. The type of networking preference of this constellation is still not clear.

The second figure illustrates the satellite constellation network with ISLs. The traffic from mobile user 1 is directly relayed from satellite 1 to satellite 2 and then satellite 3. Satellite 3 delivers the traffic down to the mobile user 2. In this type of satellite constellation, the satellites themselves with their ISLs perform a network.

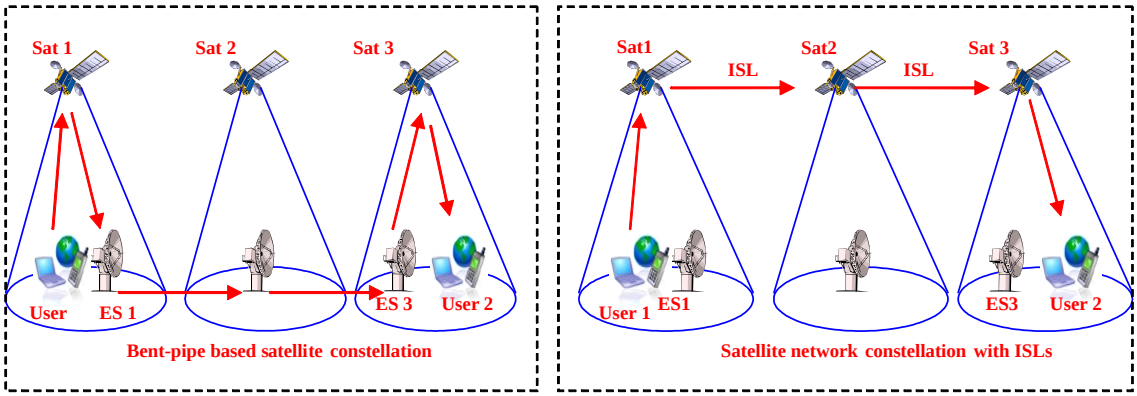

Figure 4.2.1: Satellite constellation with and without Inter Satellite Link

4.2.3 Topology of LEO Satellite Constellation

The altitude of the orbit of the LEO satellite is significant when determining the number of satellites required to provide a global coverage. Propagation delay and transmission loss decrease with the altitude of LEO satellites, but this low altitude will also decrease the coverage of a service area. The service area of a satellite is inside the coverage of this satellite footprint. At the same time, the satellite with the lower altitude will move faster, relative to the ground, to be able to stay in its orbit, which will increase the rate of handovers and Doppler Effects between terminals and satellites.

In the LEO satellite systems, a global real time service is not possible unless a complete constellation of LEO satellites is operational. The minimum number of satellites depends on the altitude and the system specifications; LEOs require a minimum of 48 to 77 satellites for

worldwide coverage, and also require that at least one satellite is always visible to each user. The number of satellites necessary to cover the whole surface of the globe (with the assumption that the satellite footprints are hexagons as in figure 3.2.4 with a central angle of $\pi/3$ *rad* and two identical angles Ψ, and has min elevation of θ_{min} and h altitude with diameter of earth as $2R$) is given in Jamalipour as follows [117]:

$$n = \frac{\pi}{3\psi - \pi} \tag{4.2.1}$$

with

$$\tan\psi = \frac{\sqrt{3}}{\cos(\cos^{-1}(\frac{R_{earth}}{R_{earth} + h_{sat}}\cos\theta_{\min}) - \theta_{\min})} \tag{4.2.2}$$

According to Akyildiz and Uzunalioglu et al. the lower altitude of the LEO satellites permits more effective communication performance with smaller and less complex user terminals. This is due to lower link attenuation [58].

Most of the proposed LEO satellite systems use circular orbits with constant altitudes and constant magnitude of circular velocity. Some other systems use multiple elliptical orbits with variable altitudes. Satellites in this elliptical system move relatively slowly around high altitude apogee, allowing users on the earth to use satellite services.As shown in figure 4.2.2, satellite 1

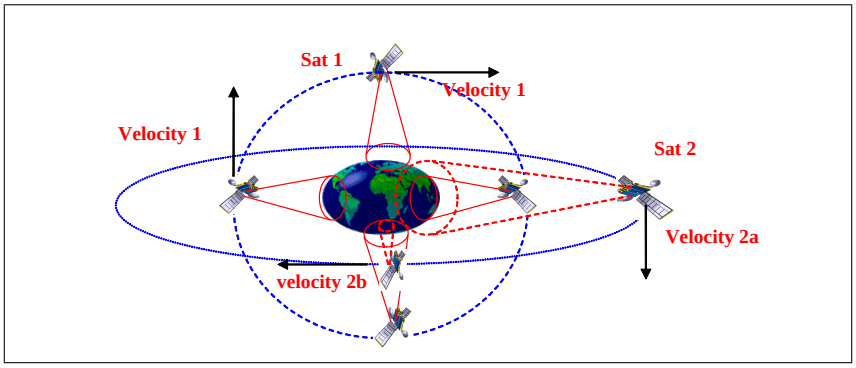

Figure 4.2.2: Different orbital shape of LEO satellite

has a constant altitude and angular velocity, velocity 1, while satellite 2 has variable altitudes and velocity. Satellite 2 has the slowest velocity, Velocity 2a, and has the largest coverage area.

At this position, satellite 2 provides service. The fastest velocity of satellite 2 is when this satellite is in the orbital position closest to earth (velocity 2b). The satellite will switch off its service at this position.

If a LEO satellite constellation has the same altitude for all of its satellites, then generally this type of satellite constellation can be divided into two categories according Wood, namely 'Walker Delta' or 'Rosette', and 'Walker Star' or 'Polar' satellite constellations. This topology is a variation of the Manhattan network as given by Wood and Pavlou et al. [32]. The first category, the Walker Delta or Rosette satellite constellation performs a fully toroidal network. In this topology, the ascending (moving in northerly direction) and the descending (moving southerly direction) planes overlap and span the full 360° of longitude. This network is used by the Hughes Spaceway NGSO. The best coverage with visibility of multiple satellites from a MT on earth can be achieved at the mid-latitudes, where the population density is high. This type of constellation network does not cover the poles. The second category, the Walker Star or Polar satellite constellation, the constellation performs a form of cylindrical mesh network. In this topology, the ascending (moving northerly direction) and the descending (moving southerly direction) planes each cover around 180° of longitude and are separated. This network is used by IRIDIUM and Teledesic. This type of satellite constellation provides a near-complete global coverage, but has an overlapping coverage at the unpopulated poles as side effect. Ground terminals will encounter an orbital seam between the last plane of ascending satellites (traveling north) and the counter rotating (or descending) satellites of the plane almost 180^0 away [118].

According to Cruishank, Sun et al. [52], an ISL between satellites in the orbital seam planes is called cross-seam ISL. IRIDIUM does not provide this type of ISL, while Teledesic allow only one cross-seam ISL in this orbital seam. In relation to these two different satellite categories, satellite constellation can be described following these two notations [52]:

1. Walker notation: this notation is given by $\{N/P/p\}$ with N is the number of satellites in one plane, P is the number of satellite planes, and p is the number of distinct phases of planes to control spacing offsets in planes.

2. Ballard notation: this notation is given by $NP/P/m$ with NP is the number of satellites in one plane, P is the number of satellite planes, and m is the harmonic factor describes

the phasing between planes.

Keller mentioned two regions in which satellites in adjacent orbital planes move in different directions. The authors called these regions counter-rotating interfaces. When satellites in adjacent orbital planes move in the same direction, it is called the co-rotating interface. In order to ensure a continuous coverage in the counter-rotating interface regions, the angle between counter-rotating planes must be smaller than angle between co-rotating planes [124].The service

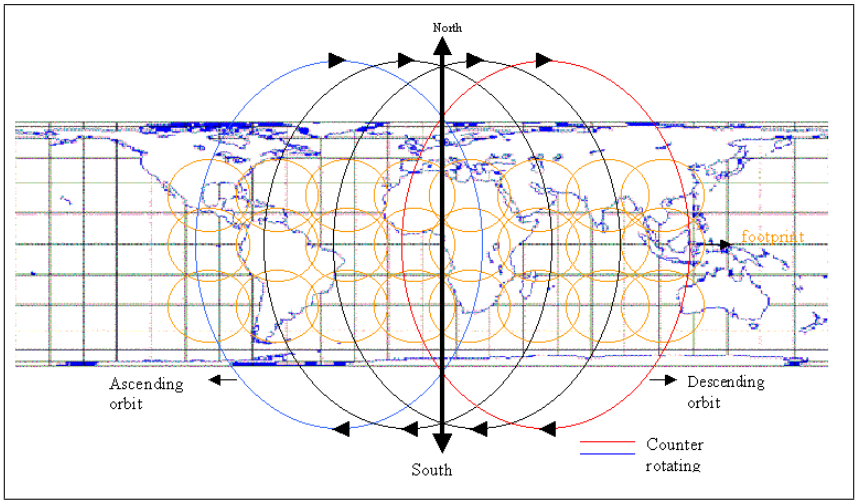

Figure 4.2.3: LEO satellite constellation footprint (background map projections is from [125])

area of a single satellite is a circular area on the earth's surface as given in (figure 4.2.3). In this area, the satellite is visible if its position in orbit is under an elevation angle equal to or greater than the minimum elevation angle determined by the system. The IRIDIUM satellite has a footprint with a diameter around 4,021 km. In order to provide a global coverage, some overlapping footprints of adjacent satellites are necessary, in which the users in these coverage areas will have more than one satellite visible (diversity). The effective footprint of a satellite usually forms a hexagon. Footprints of individual satellites are divided into smaller cells called spot beams, allowing frequency reuse inside the footprint see figure 4.2.4. Identical frequency can be reused in different spot beams (those are geographically separated) to limit interference. In IRIDIUM, each footprint consists of 48 spot beams with diameters circa 700 km [126].

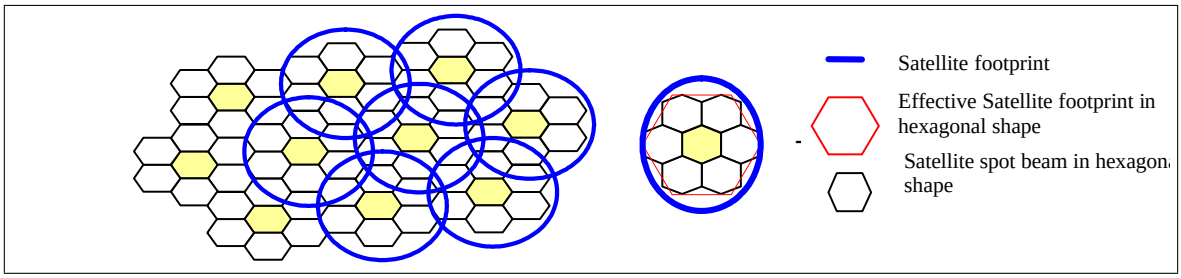

Figure 4.2.4: Satellite footprint and spot beams in a hexagonal cell form

4.2.4 ISLs and LEO Satellite's Mobility

LEO network topology is constructed by the ISL, which establish a network between satellites. In order to establish ISLs, each satellite needs to have extra equipment, including transceivers and antennas. This additional equipment will increase the weight and the cost of the satellite, but on the other hand a satellite system which has this ISL features does not require an Earth station (ES) to establish a long distance connection. This advantage reduces the dependency of the satellite network on terrestrial systems.

In the LEO satellite constellation, there are two types of ISLs. ISLs, which are permanent between satellites in the same plane (intraplane satellite link) and semi permanent between satellites in the neighboring planes (interplane satellite link) will function as edges in LEO satellite constellation with LEO satellites as the nodes. As illustrated in figure 4.2.5, ISL between satellite 1 ($sat1$)-satellite 2 ($sat2$), and ISL between satellite 1 ($sat1$)-satellite 3 ($sat3$) in orbital plane 1 are interplane ISL. The ISL between $sat1$ in orbital plane 2, $sat4$ in orbital plane 3 and the ISL between sat1 and sat5 in orbital plane 1 are intraplane ISL.IRIDIUM is the first satellite communication system that provides ISL. ISLs use radio or laser media for their direct communication. IRIDIUM uses a GSM based telephony architecture, and a geographically controlled system access process. Each satellite is connected to its four neighboring satellites through their ISLs. Connections between the IRIDIUM network and the Public Switched Telephone Network (PSTN) are provided via ES or Gateway installations. By having ISL the location of the ESs can be flexibly located.

The still under development of the Teledesic satellite system provides ISLs, which are based on connectionless packet-orientation to its eight neighboring satellites. Each satellite will act

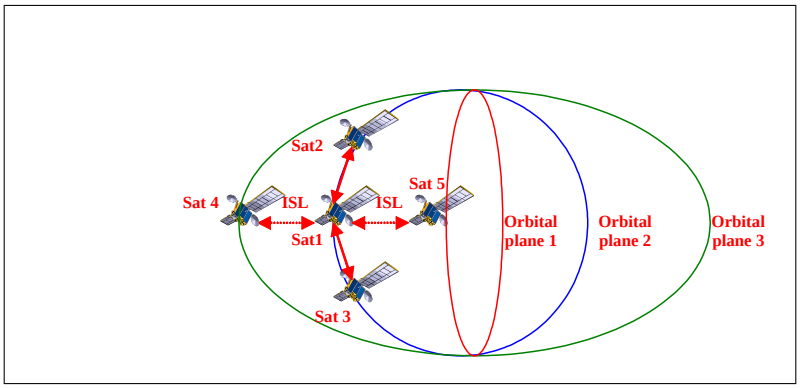

Figure 4.2.5: Satellite constellation with ISLs and satellite planes

as a switch in the mesh network of these satellite ISLs. The communication within the network uses streams of short packets with fixed length (512 bits). They use mechanisms similar to Asynchronous Transfer Mode (ATM). The capacity of each ISL is 155 Mbps. Gateways of Teledesic will provide the connection to the land fiber network, connection to the Teledesic support and database systems, connection to privately owned networks, and connection to high-rate terminals [127].

4.2.5 Mobility Management

There are different types of mobility management in the terrestrial cellular network as discussed in Chapter 3 and the mobility management in the LEO satellite network. First of all, since a LEO satellite's movement is relatively higher than the movement of any object on earth as illustrated in figure 4.2.6, a LEO satellite becomes the moving object in this context, while the mobile user on earth has a relatively fixed position. Therefore the mobility aspect that we discussed is more concerned with the movement of the satellite footprint on earth, in which an on-going connection of users on earth has to be handed over from one satellite to another. These transitions have to be smooth and seamless. Mobility management has to be able to maintain any on-going network connections, and perform a handoff process if needed. Secondly, there is an advantage in a LEO satellite's environment, in which the movement of the LEO satellite is roughly predictable. In terrestrial cellular systems, it is difficult to have a prediction of the mobility of the cellular system users.

4.2.6 Handover in LEO Satellites

Due to the rapid movement of satellites, MT users can only use the same LEO satellite for short periods of times and before they lose their LOS to the current satellite, their connection must be handed over to the next satellite. We can say that a handover is initiated from one satellite to another satellite. The term handoff is used in US cellular standard documents, while in ITU documents the term handover is used. Both terms have the same meaning.

Handover in LEO satellite systems differs from the one in the cellular terrestrial system in term of the mobile and fixed units. In the LEO satellite systems handover means a procedure of changing the assignment of a fixed unit, a Mobile Terminal (MT) on earth, from one mobile unit, a LEO satellite, to another as the LEO satellites move. While in the terrestrial cellular system a handover means a procedure of changing the assignment of a mobile unit (MT user) from one fixed unit (Base Station) to another, as the mobile unit moves as given in figure 4.2.6.The time in

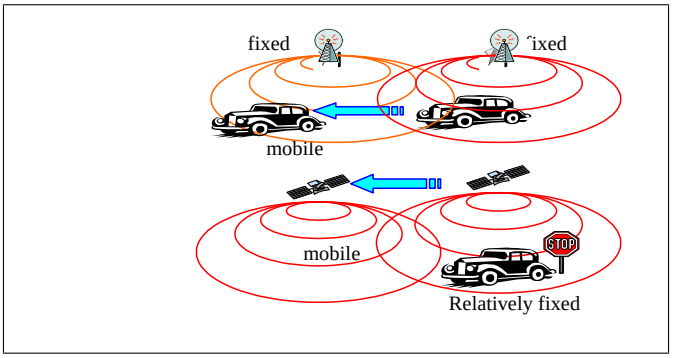

Figure 4.2.6: Different types of mobility in terrestrial cellular networks and satellite constellation networks

which a satellite is visible from a MT is called the *sliding window* of this satellite constellation. The visibility period of a satellite is defined as the maximum time duration that a MT is located inside a footprint and can directly communicate with that satellite (about 15 minutes). The satellite footprint is divided into several spot beams, and each spot beam can use several different frequencies. Because the spot beam coverage area is much smaller than the footprint coverage area, then the maximum visibility of a spot beam is around 1 to 2 minutes. This means that a spot beam handover occurs more frequently than satellite handover. A handover

can be initiated by the network, in which the decision is made by network measurements of received signals from the MT user on the ground. Alternatively, a MT user on the ground can provide a feedback to the LEO satellite concerning the signal received at the MT user. Various performance metrics may be used in making a decision, such as call dropping probability (due to the handover itself), probability of unsuccessful handover (due to an execution of a handover with an inadequate reception condition), call blocking probability (due to unavailable capacity), rate of handover (the maximum number of handovers per unit time), and handover delay (due to the distance of the point that handover should occur and the point that the handover does occur).

Stallings outlines several handover strategies [83]. Firstly he suggests a handover can be initiated by measuring relative signal strength between two LEO satellites and the MT user on the ground. The second strategy depends on threshold signal strength. A handover is initiated when the signal strength between the LEO satellite and the MT user is lower than a certain threshold value. In the third strategy, a handover can only occur if the signal strength of the new LEO satellite is stronger by a margin H than the current signal strength. The last strategy uses by using prediction techniques. The handover decision is based on the expected future value of the received signal depending on the movement of LEO satellites.

A satellite handover takes place when an ongoing connection needs to be handed over from one satellite to another satellite. There are two types of satellite handovers:

1. Intra satellite handover: handover occurs between satellites in the same plane (in figure 4.2.7 a handover of a connection of MT1 from satellite 1 to satellite 2 in plane 1). The Gateway monitors the signal strength and the portable unit's position relative to the satellite. Since the Gateway has information about the portable unit positions and the satellite positions, when the currently used satellite moves away from the portable unit, the Gateway will contact the next available satellite in the same plane, to replace the currently used satellite. The Gateway will send a message to the currently used satellite (prepare to handover the portable unit) and to the next available satellite (prepare to accept the portable unit). Then, the Gateway will send a message to the currently used satellite to let the portable unit know, and to synchronize the timing arrival of the signal.

2. Inter plane satellite handover: this type of handover occurs when a connection is handed over to another satellite in another plane, due to the unavailability of a satellite in the same plane, or due to the unavailability of satellite channels to accommodate this connection. In figure 4.2.7 a handover of MT1 connection from satellite 1 in plane 1 to satellite 3 or satellite 4 in plane 2.

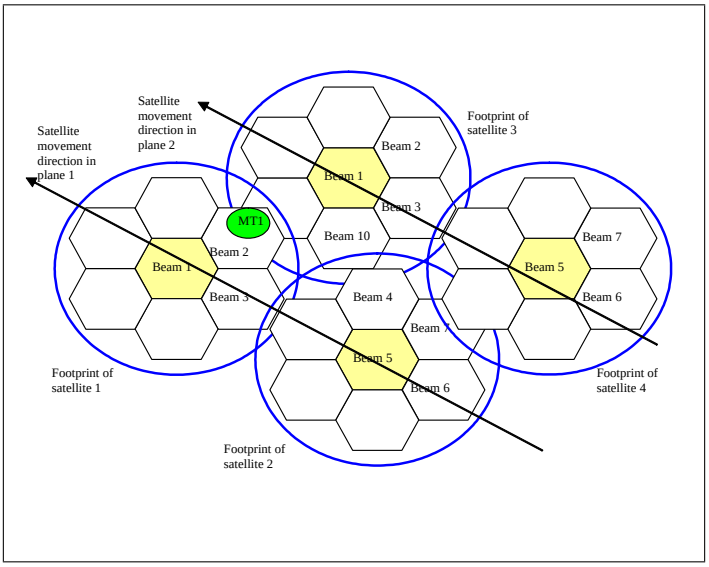

Figure 4.2.7: The movement of satellite footprints and spot beams relative to a mobile terminal (MT1) causes a handover

Stallings further identifies another type of handover: a beam handover. Beam handover occurs when a handover occurs in one satellite, and ongoing connection is handed over from one beam to another beam from the same satellite. Beam handover can be classified into two categories:

1. Intra beam handover: This type of handover occurs whilst the portable unit is still inside the same satellite's spot beam, but because availability; interference or a country's regulation, the portable unit has to use another frequency on the same beam. figure 4.2.7 illustrates a handover of MT1 connection from one frequency to another frequency in beam 2. If this happens the satellite will send a message to the portable unit to change the frequency, which means that the satellite will initiate a handover.

2. Inter beam handover: Inter beam handover occurs whilst the portable unit decides to use another frequency in the adjacent candidate beams, due to the weakness of RF signal power from the used frequency. figure 4.2.7 illustrates a handover of MT1 connection for example from beam 2 to beam 3. A portable unit monitors continuously the power strength of the RF signal in the current beam and candidate beams. Once the RF power strength in the current beam is lower than the candidate RF signal from the adjacent beam, the portable unit initiates a handover request to hand over a user to the new beam. If the handover is permitted then a new frequency will be assigned to the user. Since the same frequency can not be used in the adjacent beams, IRIDIUM uses a 12-beam reuse pattern. An inter beam handover can happen frequently; e.g. every 2 minutes or even less. In this case, the portable unit will initiate the handover.

4.2.7 Perturbations of the Satellite Orbital

Theoretically, the satellite orbit follows the two-body gravity equations, but in practice there are less than ideal factors that can invalidate this theory:

1. The earth's unsymmetrical form: The earth's diameter is longer at the equator than between the poles. This makes it more complex to define the gravity source point.

2. Solar and lunar effects: The sun and moon's gravity fields influence the earth's gravitation field in relation to the satellite.

3. Atmospheric drag: Low orbital satellites encounter the atmospheric friction of the upper layer of the earth's atmosphere.

4. Solar radiation pressure: The collision between photons radiated from the sun and the satellite cause a solar radiation pressure. This pressure is absorbed or reflected.

In our simulation model, we make an assumption that these less than ideal factors have no influence on our model. We consider only the orbital geometry, since it influences the satellite coverage and diversity.

4.3 Satellite Signal Processing in LEO satellites

4.3.1 Satellite Signals

Satellites can act as microwave repeaters echoing signals from earth stations, without any on-board processing. Tanenbaum [128] notes that communication satellites generally have up to a dozen transponders, where each one of those has a beam that covers a portion of the earth below it, ranging from a wide beam 10,000km across to a spot beam only 250km across. Transponders can have either a fixed beam to a specific earth station or a steerable beam. Earth stations within a beam region can send information to a satellite on uplink frequency. Satellite transponders will either amplify and rebroadcast them directly on downlink frequency to the destination, or amplify them and send them to the next satellite, which is closer to the destination. Different frequencies are used for the uplink and downlink paths to keep the transponders from going into oscillation and prevent interference between two links [128].

Satellites operate on microwave frequencies, between 1 to 31 GHZ as given in table 3.3.1. Microwaves signals transmitted between earth stations and satellites propagate along LOS paths and experience free space loss that changes proportional to the square of the distance, Carlson writes as [113]:

$$L = (\frac{4\pi l}{\lambda})^2 \tag{4.3.1}$$

where L is the free space loss, l is the distance in meter and λ is the wavelength in meter. In free space loss as given in (4.3.1), the path loss is considered to be ideal. There is no ground reflection or any multipath received in the receiver. This performs a minimum path loss and consider as a lower limit of the basic path loss as given in (3.3.1).

In order to transmit signals over long distances, to minimize interference over the channel and to be able to assign different channels of different frequencies modulation must take place, where the information signal to be transmitted is modulated with a high frequency carrier signal that varies some of its parameters according to the message signal. Some of the early satellite systems used modulation based on frequency modulation, but new digital modulation techniques evolved and are applied to the new satellite systems. Most digital transmissions used by satellite systems are phase shift keying (PSK) techniques, which will alter the phase of the carrier by 0^0 to 90^0 according to whether the binary signal having logic 0 or 1. Another

popular digital transmission used is the offset PSK, which gives phase shifts of 0^0, 90^0, 180^0, 270^0.

4.3.2 Signal Distortions

Schiff and Chockalingam show that as a terrestrial cellular network, LEO satellite channels are affected by random varying losses due to different signal distortions. Due to the distance between LEO satellites and the MT, the receiver will encounter transmission loss, which will represent the distance attenuation [129].

Obstacles in the propagation path such as buildings and trees cause a shadowing loss. According to Bekkers and Smits, in urban areas the shadow loss of the terrestrial cellular system typically follows a log-normal distribution which varies in the range 4dB-12dB. In LEO satellite systems the percentage of shadowed areas will continuously change with time. Shadowing is larger at low satellite elevation angels than at high satellite elevations. Especially for urban and suburban areas, shadowing will depend on the azimuth angle of the satellite as well [30].

Another type of loss, which will occur while receiving signals from LEO satellites, is called multipath fading. This type of loss is caused by differences in phases of signals received through multiple reflected paths. In cellular environments the received signal variation due to multipath follows a Rayleigh distribution.

In the cellular environment, the Doppler Effect is due to the user movement only; for typical operating frequencies (900-2000MHz) and user speeds (<100km/h) the shift in Doppler is less than 200 Hz. While in a LEO satellite constellation since the base stations (the satellites) move, even when the user remains static (e.g. fixed user terminals), there is still a Doppler due to the satellite motion. The Doppler shift due to satellite motion is relatively higher than the Doppler shift in the cellular system in the order of several tens of KHz in Globalstar L-band [25, 26].

The final cause of signal distortions is specular reflections. Since MTs antennas in the LEO system have wide-angle patterns, which tend to collect more reflected power than directive antenna, the handheld user terminal antennas may collect strong specular reflections from the ground. This will result in signal strength variation. According to Gavish eth al., variation will become higher once the surrounding area has a high reflection coefficient and as soon as satellites are in low elevation angles. This variation of signal strength causes a problem during

handovers. The strength of the signal can vary in time too, as when LEO satellite is in a low position on the horizon it will encounter a longer path through the atmosphere, which results in a higher loss in the signal strength. Gavish et al. describes more about the impact of low altitude LEO satellites into distance attenuation, Doppler Effects, power level consumption based on the number of spot beams, up/down link frequency, antenna beam openings, and the size of cells [29]. These impacts of low altitude LEO satellites into satellite signals determines the QoS that LEO satellites can guarantee to the MTs.

4.4 Switching and Routing Processing

Depending on the type of transmitted signal and the expectation QoS, mobile satellite communication channels can provide three methods of channel processing:

1. Store and forward packet data channels are used to transmit small amounts of user data with delivery times of several minutes. This type of channel is typically used for cargo tracking services, paging and some emergency distress signaling.

2. Interactive packet data channels are used for services when a several minute transmission delay is unacceptable. These services are typically interactive messaging services.

3. Circuit switched channels are used for applications requiring real-time voice communications or for transmitting large amounts of data, such as facsimile or file transfers.

In some satellite constellations, routing of incoming calls from source to destination is performed in Gateway stations (BP-Sat). Recently, routing processing has begun to be performed on the satellite itself (XC-Sat and SW-Sat). Once there is an incoming call, the satellite decides which path a current call has to follow. IRIDIUM is the first satellite constellation to provide this onboard processing. A call from a MT can be routed within a satellite network and connected to any MT in any location, or it can be connected to a public network through any ES. The IRIDIUM system is based on GSM call processing architecture (see chapter 3), and ESs will be connected to a GSM Mobile Switching Centre (MSC) with associated databases: Equipment Identity Register (EIR), Home Location Register (HLR), and Visitor Location Register (VLR). Some additional functions that are special to IRIDIUM system and not to GSM MSC should be

taken care of by ES. When a MT originates a call, the IRIDIUM system will calculate the user's location. Each ES has information of the location area that the ES controls. MT locations will be used to assign home ES (or visit ES, if the MT has roamed) which will control all procedures for making a call. Based upon the MTs location and information about PSTN/PLMN (Public Land Mobile Network) at that location, a PSTN/PLMN can be connected by ES to set up the call. Using the MT position also, ES ensures compliance with national laws enforcing call restrictions on MTs. The use of ISLs will remove the requirement for the ES to be continuously available within the satellite footprint. Also, by using ISLs terrestrial charges can be kept to a minimum, by routing a call to use ISL as far as possible to the closest ES to the origin, and destination of the particular call. IRIDIUM uses a mixture of Time Division Multiple Access (TDMA) and Frequency Division Multiple Access (FDMA) for its multiple accesses.

Teledesic provides onboard processing as well. Each satellite that serves an associated cell manages channel resources (frequencies and timeslots) that can be used by this cell. A MT uses the same channel resources during a call; irrespective of which and how many satellites are serving this MT. Using this method channel reassignments would only happen by the handover. A database onboard each satellite is used to avoid interference between cell areas. Teledesic choose a combination of multiple access methods: space-time-and frequency-division multiple access. Each super cell associates with one beam. Each spot beam is divided into some number of cells, which can use the same spot beam. At any time only one cell will be scanned by this spot beam. This is the TDMA part of the Teledesic multiple access method. Multiple access between cells in a spot beam coverage area is called supercell. The Space Division Multiple Access (SDMA) is used between cells scanned simultaneously in adjacent supercells. Transmissions from satellites are synchronized so that each supercell will receive transmissions at the same time. In order to ensure that there is no overlap between signals from cells a guard band percell is used, while the FDMA part is in the cell's time slot. Each terminal makes use of FDMA in the cell's time slot for uplink and ATDMA (asynchronous TDMA) for downlink. Each terminal is allocated into one or more frequency slots on uplink for the call's duration, and on downlink a 512-bit packet header to separate users (rather than using a fixed assignment) is used. Due to space separation of the Teledesic system, all supercells can use all available frequencies. However, only one of the nine cells in a supercell uses all available frequencies at

one time [127].

A connection request from one MT on earth will be served by a LEO satellite by providing an available route in LEO satellite network from source to destination. A routing procedure will decide which route a new connection should be allocated. Wood studied a simulation of delay time for a traffic sent from a ground terminal in London, England to Quito, Ecuador for different types of satellite constellation The Teledesic proposed routing procedure achieved the shortest delay path compared to a hop via two GEO satellites, a hop via a single GEO satellite, and a Spaceway NGSO proposed routing procedure [118]. In each of these routing procedures certain routing algorithms are included, which are needed to determine the best way to route the incoming traffic.

Communication companies in the world submitted their proposals for various satellite constellations to the FCC in order to acquire their licenses. In their proposals different types of onboard switching are suggested. Some companies consider an ATM-based switch as their onboard switching such as the GEO based Spaceway satellite system. Some other companies consider packet switching as their onboard processing for their LEO satellite system such as the Teledesic system with its own designed protocols over ISLs and in the earth space interface [95].

4.4.1 Satellite Network Protocol

To some extent, satellite network characteristics will be different from those in terrestrial networks. A network interface has to be deployed before these two networks can interact with each other. The satellite network interface unit, which is normally located in the Gateway will convert and map terrestrial network protocol to a satellite network protocol and vice versa. The characteristic difference between satellite and terrestrial networks is found mainly in the physical layer when an end to end or point to multipoint communication has to be accomplished.

Research by Emmelmann, Brandt et al. [63], made an assumption that the future satellite systems are based on ATM network infrastructure, in which we can consider each satellite as an ATM node, each ISL as a single Virtual Channel Connection (VCC), and the routing path of a connection as a Virtual Path Connection (VPC). The system concept is given as in the figure 4.4.1 [63]:In figure 4.4.1 a traffic request from a user on the earth is converted by an ATM Adaptation Layer (AAL) into an ATM traffic format with their different type of classes

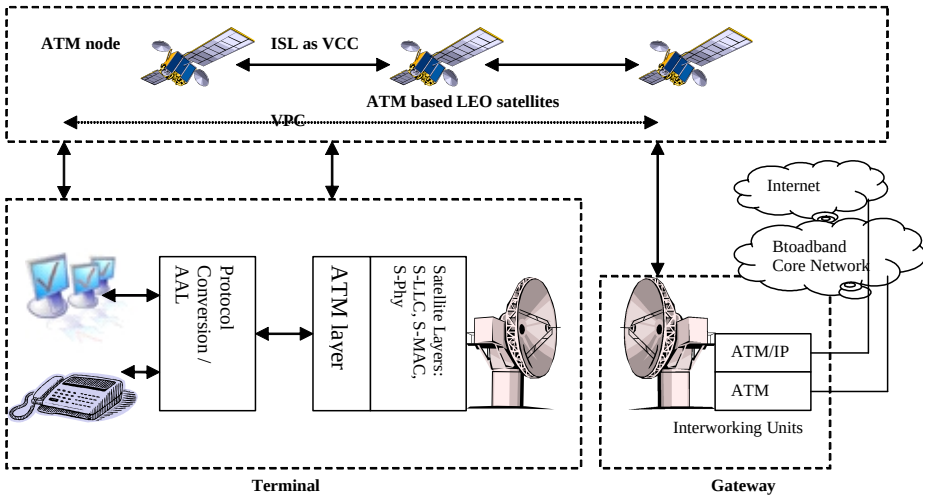

Figure 4.4.1: ATM based LEO satellite network

(CBR,UBR, ABR, and VBR). It goes through the ATM layer and will be transmitted to the LEO satellite by the satellite modem, which provides the satellite communication layers. Traffic from other users, including Internet traffic and broadband traffic from the IP network or other broadband core networks, goes through the Interworking Units (IWU) before being transmitted by the satellite modem in the Gateway to the LEO satellites.

With the increased use of the internet, more researchers are looking for the implementation of IP technology into satellite networks. Ekici, Akyildiz et al. propose a network layer integration between terrestrial and satellite IP networks. The communication between these two networks is to be enabled by introducing a new exterior Gateway protocol called Border Gateway Protocol-Satellite Version (BGP-S). The hybrid terrestrial, satellite network architecture is given in figure 4.4.2. In this proposal, the satellite network is considered as a separate Autonomous System (AS) with a different addressing scheme. The terrestrial Gateways act as border Gateways on behalf of the satellite network and perform the conversion of addresses. Then an exterior Gateway protocol such as Border Gateway Protocol (BGP) can find a path over both networks. BGP-S will support the automated discovery of paths that include the satellite hops. The terrestrial internet is presented as ASs, with its Interior Border Gateway Protocols (IBGPs), Exterior Border Gateway Protocols (EBGPs). Active Peer Register (APR)

provides a list of active peer Gateways connected to the satellite network [76].The interface

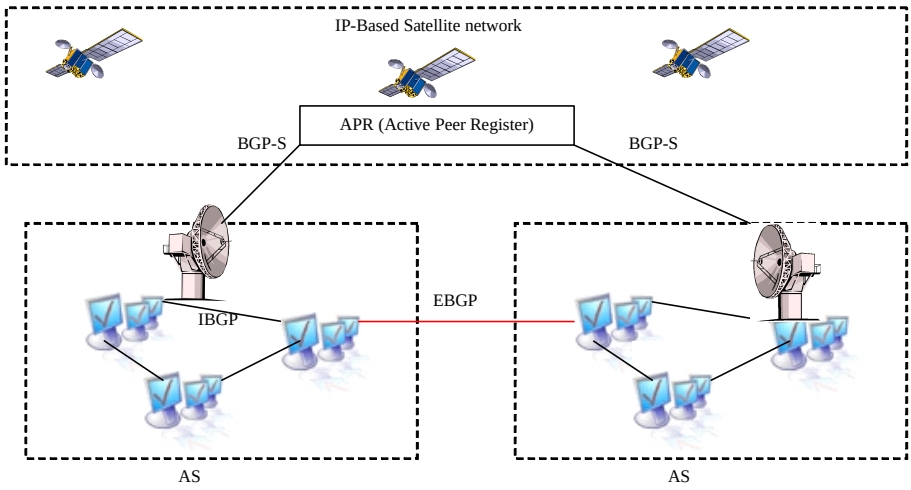

Figure 4.4.2: IP-based LEO satellite network

between the terrestrial and the satellite network was introduced in both papers of Ekici, Akyildiz et al. and Emmelmann and Bischl. The satellite network interface unit performs several tasks to interface satellite network protocol and terrestrial network protocol. A protocol mapping or tunneling between the terrestrial network protocol and the satellite network protocol is necessarily performed, in order to achieve communication between the satellite and terrestrial networks [76].

They further show that the satellite interface network unit performs a multiplexing and demultiplexing of the uplink and downlink streams from the satellite channel. Shaping is introduced in the Satellite Interface Network unit (SIU), which can be used to preserve characteristics of packets at the entrance of a satellite network. Introducing this shaping buffer can improve throughput of the satellite network [64].

If we consider networking layers of different types of satellite constellations, (BP-satellites and the XC-satellites/SW-satellites) the corresponding routing approach differs only in the space segment, satellite constellation. The ground segment of those satellites, Satellite Network Interface Unit (SIU), has an almost similar structure. In XC and SW, there is a connection between the Medium Access Control (MAC), Logical Link Control (LLC) and physical layer

of one satellite into another satellite. A BP-satellite control only uses the MAC/LLC and the physical layer of one satellite. In figure 4.4.3 and figure 4.4.4 the protocol stack of both BP-satellite and XC/SW satellites is shown.In some XC/SW-satellite constellations the loaded

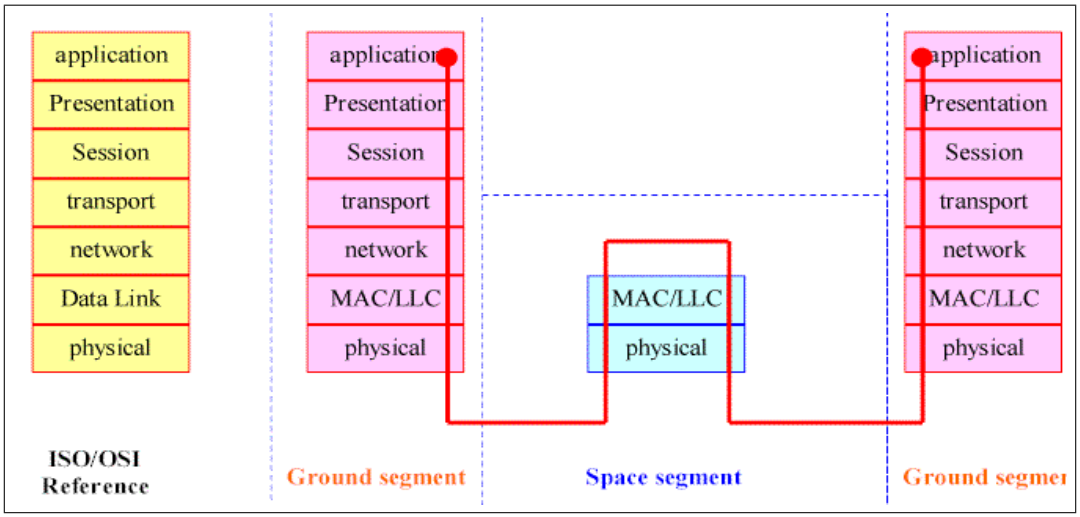

Figure 4.4.3: Network Layers of a Bent-Pipe Satellite system and the corresponding ISO/OSI reference model [52]

traffic is considered as multiplexed streams of short fixed-length packets, which contain headers. Conversion of the packet format from the terrestrial network protocol to the satellite protocol will take place in the Gateway, before the packet is forwarded up to the satellite. The Gateway will also be responsible for the multiplexing and demultiplexing of the streams from the MTs.

In the satellite constellation, the MAC/LLC layer becomes significant since it uses multiple access to share the available communication medium. MAC/LLC will function as Radio Link Protocol (RLP) in the terrestrial wireless environment (see chapter 2).

The main multiple access schemes in MAC/LLC are FDMA, TDMA, and Code Division Multiple Access (CDMA). Detailed comparison between these three types of multiple access scheme is given by Lutz and Stallings [53,94]. In addition the MAC layer scheme can be classified into three different access methods, which are suitable for traffic with QoS requirements as put forward by Cruickshank, Sun et al. [52]. According to Cruickshank, Shun et al., MAC layer schemes can be classified into the fixed access method, Demand Assignment Multiple Access (DAMA), and adaptive access.

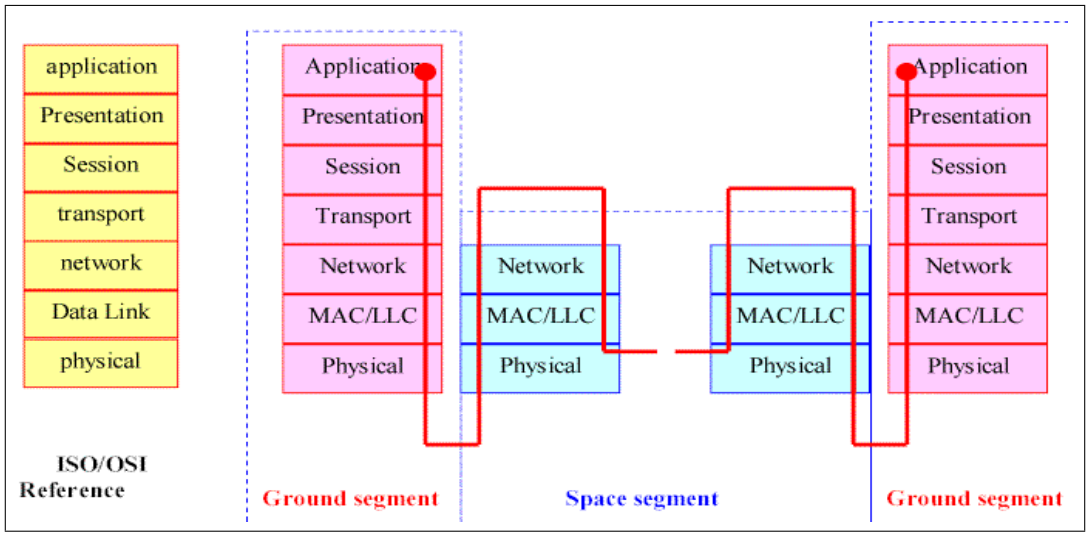

Figure 4.4.4: Network Layers of SW/XC satellite constellations and the ISO/OSI reference model [52]

1. The fixed access method is used in order to meet the needs of constant high traffic of long duration. Available slots are allocated for specific terminals during the lifetime of the terminals. If they do not occupy the allocated slot, then the slot will be wasted. Lutz introduces channel borrowing scheme to increase traffic allocation efficiency. In this scheme, if all channels in a satellite spot beam are fully occupied then a channel from a neighboring spot beam is made [53]. According to Re, Fantacci et al., the number of channels permanently assigned to each cell depends on the number of system resources and the reuse factor [60].

2. The Demand Assignment Multiple Access method is used in order to meet the needs of sporadic traffic of long duration. There are two types of DAMA - fixed rate and variable rate. Fixed rate DAMA has almost similar properties to fixed access scheme but a slot is allocated to the terminal only during the connection, not during the lifetime of the terminal. Variable rate DAMA allocate slots only when there are cells waiting for the service. This will reduce the rate of collision. The drawback is corresponding delay, which will start when the terminals ask for a slot and continue until the MT actually can transmit their cells. According to Chitre and Yegenoglu, DAMA is implemented in the COMSAT Linkway 2000 [130].

3. The adaptive access method is used in order to meet the needs of multiple media with different characteristics. Random-Reservation Adaptive Assignment (RRAA) is one of the variations in adaptive access, which combines the three different types of accessing method. RRAA is used in both LEO and GEO satellites.

Another type of access method, which it is not suitable for traffic with QoS requirements, is called Random Access method. Random Access method is used in order to meet the needs of constant high traffic with short to medium durations. Favorite Random Access is Aloha, this access method allows MTs to transmit simultaneously. If a collision occurs then a retransmission is necessary. This type of multiple access is not suitable for QoS, but well suited to best effort service.

Once a traffic demand has an access to a satellite constellation, then the satellite networks need to allocate a certain route from source to destination for this incoming traffic. Since the increased use of the internet in global communication, IP routing has become an interesting research topic. Wood, Clerget et al., discuss three approaches to provides an IP routing interface between terrestrial and satellite networks. The first approach is tunneling. The tunneling approach can be used to isolate the constellation network's routing from external networks (terrestrial networks). A satellite constellation performs as a autonomous system and can create tunnels across the network that link IP-capable entities on the ground. In order to be able to send traffic from one edge of the tunnel to another, it is necessary to know the constellation address. These addresses are provided by a Constellation Address Resolution Server (C-ARS). There are some disadvantages of this tunneling system because tunneling requires a processing overhead, and it can give an unfair picture of the number of hops between two points. All tunnels appear to have the same length, a single IP virtual hop. Since the header is encapsulated, the Time to Live (TTL) hop-count is not decremented in the tunnel. The second type of approach is using Network Address Translation (NAT), which translates the internal address realm in the IP packet into a suitable external address in the other realm. The satellite constellation network can be viewed as a private network. NAT introduces some implementation problems, such as the renumbering of adding an extra ISL. The last approach is by using an external routing for constellation networks, such as adding a Border Gateway Protocol (BGP), with their satellite version as mentioned before in this chapter S-BGP [72].

4.4.2 Signal Blocking and Satellite Buffers

In order to increase the bandwidth availability, a higher frequency band is used. However, a higher frequency in general has a greater effect of transmission impairments because of attenuation loss due to atmospheric changes and the long distance between antenna and satellite antenna. For a given frequency allocation for a service, a higher frequency is allocated for a downlink and an uplink band. The higher frequency suffers from greater free space loss (greater spreading) than its lower frequency. This loss results in a weaker signal received by a satellite or a MT on earth. Due to this weak signal, sometimes during a connection between a satellite and the MT on earth, the satellite/MT cannot maintain the minimum required signal strength for a certain period of time. This results in a dropping of the traffic channel to the MT.

In addition, the round trip delay time in satellite communication is relatively high compared to the transmission time of a single frame. In case an error occurs in the transmission, the general approach is to retransmit the error frame and all subsequent frames. In satellite communication with a long data link, an error in a single frame will require retransmitting many frames, which can result in the blocking of other incoming frames. Therefore, according to Stallings it is likely to enable the receiver to correct errors in an incoming transmission on the basis of the bits in that transmission. One example of this method is Forward Error Correction (FEC) [83].

Signal blocking occurs in two different situations. When a MT initiates a new connection via a satellite, the corresponding satellite allocates a traffic channel for this request, if there is an available capacity left. If there is no available capacity after several attempts, the connection will be blocked (call blocking). The second situation occurs when a handover has to be initiated from one LEO satellite to another. An existing connection is dropped in the handover process, because of an overloaded capacity in the new LEO satellite (call dropping).

Both situations occur because the number of channels available in the corresponding LEO satellite is less than the number of potential users, who request the LEO satellite service. For a blocking system, the fundamental performance parameter is the blocking probability that a request will be blocked.

In the case that a blocking system has a buffering scheme, then the 'not yet' serviced by LEO satellites will be put in a queue. The next fundamental performance parameter is the

average delay of the waiting period.

The difference between the satellite environment and the terrestrial wireless environment can be distinguished by their buffering scheme. Cruickshank, Sun et al., noted that in the terrestrial network, MTs can transmit at their peak bit rate and will be buffered in their buffer system. Bandwidth constraint will be activated on the outgoing link after buffering. While in the satellite constellation, bandwidth constraint will be activated before the buffering/multiplexing point. In this scheme, the transponder's available power and bandwidth are shared between MTs and Gateways, before buffering [52].Figure 4.4.5 shows that incoming traffic into the

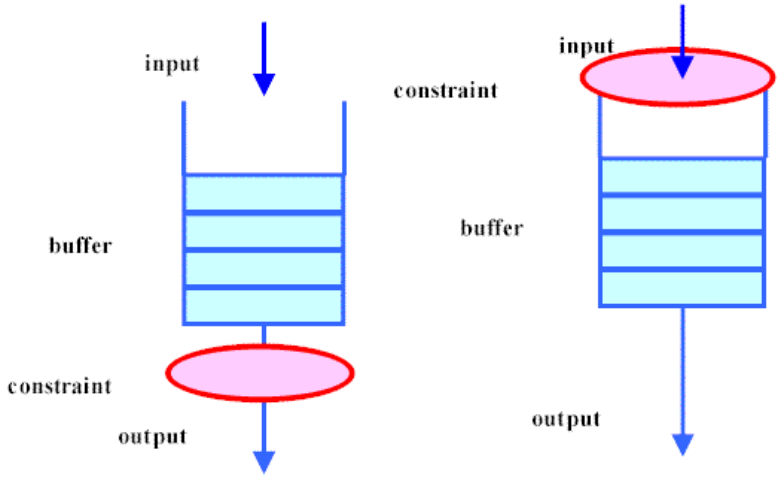

Figure 4.4.5: Buffering system in terrestrial network and satellite network

terrestrial network is buffered following the peak rate of each MT, and bandwidth constraint is performed as soon as the traffic flows out of the stations. While in satellite communication, bandwidth constraint is already being processed to satisfy the constraints before the satellite's buffers receive it. Therefore, each MT can only transmit following the bandwidth constraint of the satellite communication.

The blocking system can be defined by two factors: the manner in which blocked calls are handled and the number of traffic sources. A blocked request for a connection can be handled in two ways according to Stallings. First the blocked request will be put in a queue and wait for a free channel which is referred to as Lost Calls Delayed (LCD); otherwise the blocked request will be rejected and dropped. The second method leaves two assumptions about the action of

the user. The first assumption is that the user hangs up and waits some random time interval before attempting to make another call. This is called Lost Calls Cleared (LCC). The second assumption is when the user repeatedly attempts calling. This is known as Lost Calls Held (LCH) [83]. Kobayashi, Yu et al., proposed various blocking system, which is called a loss system [21, 131]. In our studies, we introduce satellite network as a loss network.

4.5 Summary

The differences between GEO satellites and LEO satellites show the difficulties and disadvantages arise when dealing with a dynamic topology of LEO satellites, in contrast to their advantages in reducing the significant propagation delay. There are different types of LEO satellite systems, as in IRIDIUM satellite system, which uses ISLs to complete a connection, and the GLOBALSTAR satellite system, which uses a BP-architecture. In this thesis, an IRIDIUM-like satellite constellation with ISLs between the satellites is considered, because the ISLs perform a kind of network by itself in the space, and this type of satellite system is more independent than the terrestrial system. However, a consideration of the terrestrial network system and architecture has been discussed, in order to integrate the satellite network and the terrestrial system. Various ways of interfacing both systems are given such as the tunneling method, translation of network address, and using an external Border Gateway Protocol. Issues arising from the dynamic properties of the LEO satellite, such as the different type of handovers are discussed and the satellite handover is considered including the interplane and intraplane handover. Finally, two performance parameters for the evaluation of LEO satellite systems are given, the call blocking probability and the delay time.

Chapter 5

PROBLEM FORMULATION

5.1 Introduction

In this chapter, a problem formulation and the methodology to solve the traffic allocation problem in a dynamic LEO topology are discussed. First, a description of the problem in traffic allocation for multiservice traffic in the LEO satellite network is given. Thereafter, a description of the LEO satellite system, which is the focus of this thesis is given and includes the dynamic topology of the LEO satellites and the type of satellite cells. Issues regarding handover and ISLs follow.

5.2 Dynamic Topology of a LEO Satellite Constellation

The popularity of the internet and multimedia applications, in either a fixed network or mobile network, using wired or wireless technology, has increased the instability in the amount of traffic loaded into the links, for a particular time interval. High variations of traffic load in the network links effect the complexity of routing protocol. Different applications require different quality of service, which attach added complexity to this routing problem.

This complexity is due to the type of traffic, which needs to be transferred through a satellite network. Furthermore, the LEO satellite network by itself introduces an extra difficulty to provide optimal traffic control. Since LEO satellites are moving continuously, network topology becomes dynamic. There are two different approaches to describe this dynamic topology, based

on satellite cells. Each satellite will enclose several installed spot beams and each spot beam will contain several cells. According to Henderson and Katz, we can consider the cells as a satellite fixed cell or as an earth fixed cell. In a satellite fixed cell, the satellite is considered as having a fixed, non-steerable spot beam. The corresponding satellite's footprint will move relative to a particular position on the earth. In contrast, we can consider the cell as an earth fixed cell. In this case, the satellite contains a steerable spot beam. The satellite spot beam maintains a fixed satellite footprint on the earth. The satellite's spot beam is aimed at a fixed region on earth. Therefore, a region on earth could be serviced by the same satellite for a longer period [24]. The distinction between these two approaches is explained in figure 5.2.1 and 5.2.2: Figure 5.2.1 shows three satellites attached with non-steerable spot beams (satellite fixed cell).

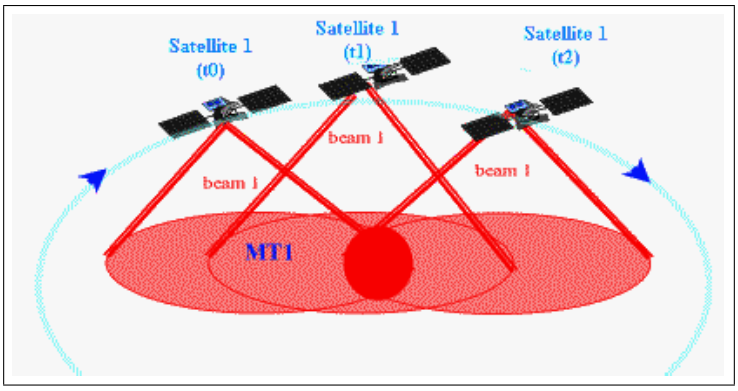

Figure 5.2.1: Satellite fixed cells

*Satellite*1 travels from time t_0 to time t_2 following clock direction. An observer, MT1 is able to make use of the service coverage of *satellite*1 from time t_0 to t_1. At time t_2, *satellite*1's footprint fades away from MT1. All existing connection from MT1 needs to be handed over to the next available satellite before t_2.

Figure 5.2.2 shows that *satellite*1 has the ability to steer the spot beam and maintain fixed cell coverage on earth (earth fixed cell). *Satellite*1 starts to provide service for MT1 at time t_0. At this time, MT1 is approached by *satellite*1's fixed cell coverage (satellite fixed cell). Nevertheless, at time t_2, *satellite*1 can provide MT1 with a service, but the service begins to fade away. Therefore, an open connection of MT1 can utilize the same satellite during t_0 to t_2.Consequently, an earth fixed cell based satellite requires less handover mechanism than a

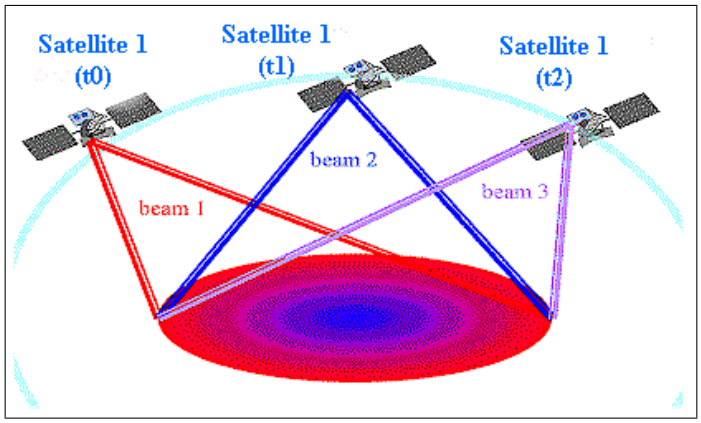

Figure 5.2.2: Earth fixed cells

fixed cell satellite. This leads to a less complex handover procedure in earth fixed cell compare to satellite fix cells. Due to this complexity reduction, our proposed model will be based on this earth fixed cell model.

In order to determine a LEO satellite position at a certain time, we need to know the orbital shape and the altitude of the LEO satellite system. Let us consider a circular orbit. In a circular orbit, a LEO satellite moves in its orbit in a relatively constant magnitude angular velocity with a direction perpendicular to the radial direction. The orbital period follows from (3.3.2) and (3.3.3). The orbital period is derived when the centripetal force is equal to the gravitational force:

$$T = \frac{2\pi R_{sat}^{3/2}}{\sqrt{g m_{earth}}} \tag{5.2.1}$$

Where R_{sat} is the total height of the satellite from the centre of the earth in meter (a summation of altitude of the satellite h and the earth radius R), g is the earth's gravitational force and m_{earth} is the earth mass. In a circular orbit, the angular velocity of this LEO satellite is then given as follow:

$$v_{circular} = \sqrt{\frac{g.m_{earth}}{R_{sat}}} \tag{5.2.2}$$

In the case of an elliptical orbit, the magnitude of angular velocity varies. Minimum velocity

occurs in apoapsis and the maximum velocity occurs in periapsis (see figure 5.2.3).

$$v_{\min}(inapoapsis) = \sqrt{\frac{g.m_{earth}(1-e)}{R_{sat}(1+e)}} \tag{5.2.3}$$

$$v_{\max}(inperiapsis) = \sqrt{\frac{g.m_{earth}(1+e)}{R_{sat}(1-e)}} \tag{5.2.4}$$

where e is the eccentricity of the elliptical orbit, (see figure 5.2.3)

$$e = \frac{R_{apoapsis} - R_{periapsis}}{R_{apoapsis} + R_{periapsis}} \tag{5.2.5}$$

Figure 5.2.4 shows the circular speed of both circular and elliptical orbits. The circular orbit

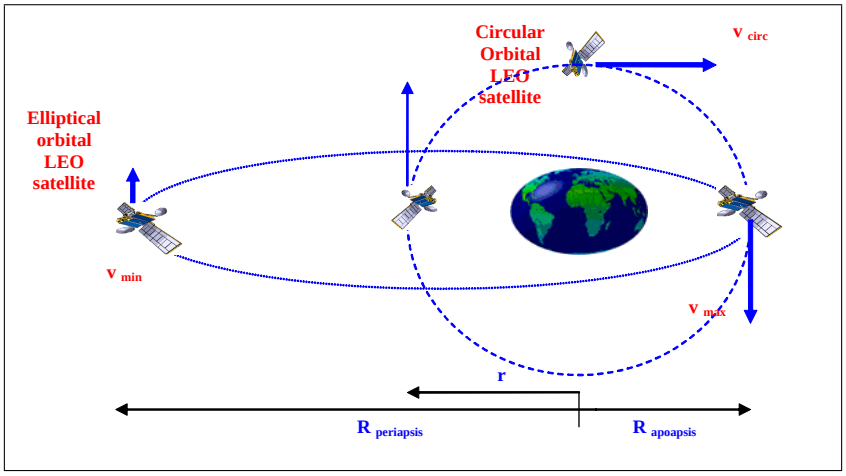

Figure 5.2.3: LEO satellite position with their corresponding angular velocity in circular and elliptical orbit

has an altitude of 700 km. In the circular orbit, the circular speed remains constant. Meanwhile in the other two elliptical orbits, the circular speed varies with their longitude positions. The first elliptical orbit has an altitude for the apogee of 700km and perigee of 1500km. The second elliptical orbit is from the Molnya Russian satellite orbital. The altitude of its apogee is about 40000km and the altitude of its perigee is about 1000km. It makes this type of satellite orbit useful when the satellite is around its apogee. The time of this satellite orbit is about 12 hours.We consider in this thesis, a LEO satellite constellation with a circular orbital, with

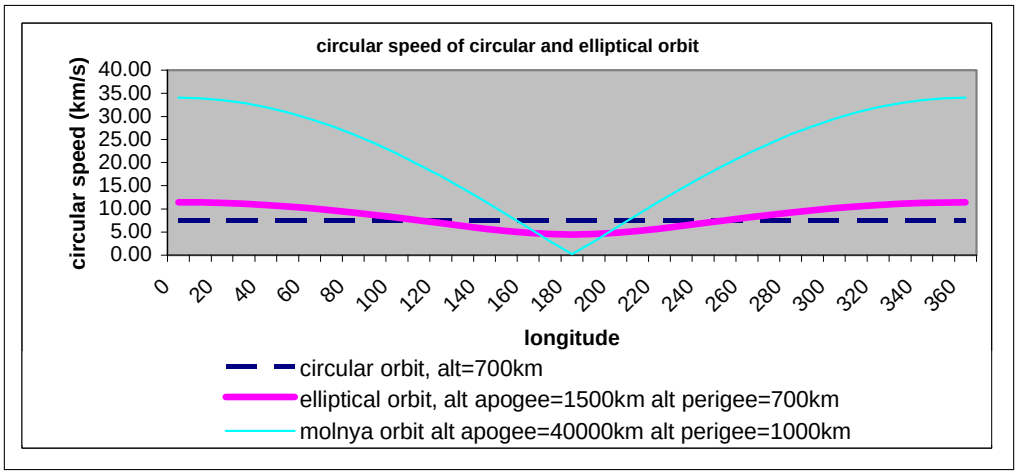

Figure 5.2.4: Circular speed of a circular and two elliptical orbits

constant magnitude angular velocity. Since the magnitude of angular velocity is constant, which can be derived from the latitude of its LEO satellite constellation, the LEO satellite position in a certain time period, can be predicted. This prediction of satellite position is valid since we consider an ideal situation in our LEO satellite constellation, wherein, no perturbation due to the atmospheric layer occurs. This dynamic but predictable characteristic of the LEO satellite delivers a good way of solving the handover problem, which will enhance the performance of our routing algorithm.

5.3 Problem Formulation

In this section, we define the problem that we are facing in traffic allocation within the LEO satellite network. First, we consider the routing problem. In this routing problem, we are looking for an available path from source to destination. A path is available only if there is a sufficient bandwidth remaining in each link of this path to accommodate the required incoming traffic. The total delay time from source to destination has to be less than the permissible delay time of the incoming traffic. Other constraints that we include are blocking probability and the node degree constraint. In addition to this routing problem, there is an extra problem cause by the dynamic property of LEO satellite topology. We consider this problem later in this chapter. Moreover, we need to consider different QoS requirements of various types of traffic. In our case,

we consider only two types of traffic, high and low priority of traffic. The remaining bandwidth of each link is demanded by these two traffic classes. In order to satisfy their required QoS, we add a privilege parameter in our routing algorithm. We introduce privilege parameter in chapter 7. This parameter defines the reserved bandwidth for high priority traffic and the maximum length of diverted path for low priority traffic. We try to optimize the routing problem to allocate various types of traffic with different QoS requirements. This optimization problem has an objective to achieve a minimum cost by considering delay time with the constraints, and to maximize the remaining bandwidth by considering the two traffic types.

In a satellite communication system along with ISL, connection will probably be composed of more than one link (see figure 5.3.1):

$$MT_1MT_2 = MT_1S_k + S_{k,i} + S_{i,\ldots} + S_{\ldots,l} + S_lMT_2 \tag{5.3.1}$$

where:

MT_1MT_2 = connection between Mobile Terminal *1* and Mobile Terminal *2*

MT_1S_k = up link connection from Mobile Terminal *1* to Satellite k

$S_k, S_i, S_l, S_{\ldots}$ = Satellite k, satellite i, satellite l, and satellite $\ldots$.

$S_{k,i}$ = inter satellite link between satellite k and satellite i.

S_lMT_2 = down link connection from Satellite l to Mobile Terminal *2*The related

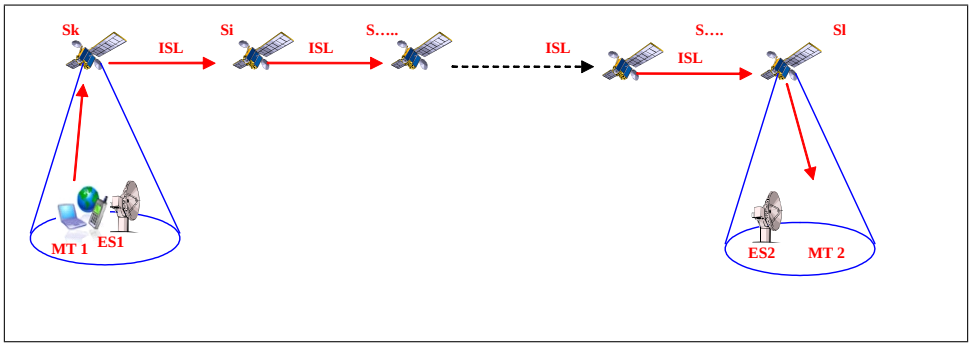

Figure 5.3.1: Satellite connection from Mobile Terminal 1 to Mobile Terminal 2

propagation delay from MT1 to MT2 (t_{prop}) will be:

$$tprop = t_{uplink} + \sum_{l \in P} t_{ISL'sprop,l} + t_{downlink} \tag{5.3.2}$$

with t_{uplink} , $t_{downlink}$ (in sec) the propagation delay of up- and down-link channel respectively. $t_{ISL'sprop}$ is the total propagation delay in the ISLs (l) belongs to path P.

Propagation delay in ISLs depends on whether the ISL is between satellites in the same plane (intraplane), or between satellites in the neighboring planes (interplane). Ekici, Akyildiz et al. show in their paper [69] that length of ISL (ISL_{length}) in the intraplane, is constant:

$$ISL_{length} = \sqrt{2}(R_{earth} + h_{sat})\sqrt{1 - \cos(\frac{2\pi}{N/n_{planes}})} \tag{5.3.3}$$

with R_{earth} is the radius of earth, h_{sat} is the altitude of satellite, N is number of satellites in the satellite constellation, and n_{planes} is the number of planes in the satellite constellation. The length of ISL in interplane is variable and depends on the latitude of the satellite, θ:

$$ISL_{length} = \sqrt{2}(R_{earth} + h_{sat})\sqrt{1 - \cos(\frac{2\pi}{2 \times n_{planes}})} \times \cos\theta \tag{5.3.4}$$

The uplink delay is also dependent on the uplink access mechanism and resource allocation approaches: Random access, fixed assignment, fixed-rate demand assignment, variable rate demand assignment, free assignment. These approaches are explained in more detail in [54–56].The processing time in a satellite, t_{proc} plays a role in the total delay of the end to end connection. t_{proc} is processing time delay due to processing time to allocate a channel in order to satisfy traffic demand. Total delay t_{delay} in ISLs between source to destination is

$$t_{delay} \leq t_{prop} + t_{proc} \tag{5.3.5}$$

The processing delay depends on the type of processing that has to be undertaken in order to allocate a flow.

Guerin and Orda in their paper [9], give the processing delay for an end to end delay bound

for a link as

$$t_{proc} = \frac{B}{r_{\min}} + \frac{\sum_{(k,l)\in p} v_{k,l}}{r_{\min}} \tag{5.3.6}$$

B is the size of the flow bursts in bits, r_{min} is the minimal rate that can be guaranteed for end to end connections in bit/second, (k, l) is the link between satellite k and satellite l in the path p, and $v_{k,l,max}$ is the flow's maximum packet size in the link (k, l) in bits.

As in (5.3.5) total delay is composed of the propagation delay, which is subject to the number of hop-lengths, distance between satellites, altitude of a satellite's orbit, and the processing delay, which is subject to the processing rate in each satellite and available bandwidth. We are interested in implementing a routing control which will optimize the performance of the satellite network, in the sense of total delay. We consider the delay as a cost in our optimization problem.

The formulation of our optimization problem is:

$$\min \sum_{k=1}^{n} \sum_{l=1}^{n} (z_{total,kl} \nu_{kl}) \tag{5.3.7}$$

$z_{total,kl}$ is the total cost function of all connection using the ISL between satellite k and satellite l in order to satisfy satellite network traffic demand between satellite k and satellite l, ν_{kl} as given below.$\nu_{i,kl}$ is the traffic load between satellite k and satellite l, which belongs to the connection demand i

$$\nu_{kl} = \sum_{i} \nu_{i,kl} \tag{5.3.8}$$

From (5.3.7), the total cost function $z_{total,kl}$ is composed of $z_{prop,kl}$, which is due to the propagation delay, and $z_{proc,kl}$ which is due to the processing delay :

$$\min(\sum_{k=1}^{n} \sum_{l=1}^{n} ((z_{prop,kl} + z_{proc,kl})\nu_{kl})) \tag{5.3.9}$$

The optimization of the satellite network's performance based on various QoS requirements is considered. Suppose that there is C traffic classes with different QoS requirement D. If only the delay parameter as the QoS parameter is considered, then D_c will represent the maximum acceptable delay requirement of application in class C.

A *path* P can be taken into account to satisfy a class C traffic demand if its total end to

end delay, $t_{total}(P) \leq D_c$. We assume in our case, that all propagation delays from neighboring satellites using ISLs are identical. This means that the propagation delay of an ISL using link (k, l), $t_{ISL'sprop,l}$ will be converted into $t_{ISL'sprop}$. The residual rate of each link (k, l), $r_{k,l}$ will have different values subject to the available bandwidth on that link, $b_{k,l}$. as:

$$r_{k,l} \triangleq b_{k,l} \tag{5.3.10}$$

We also assume an exponential distribution is used to model the distribution of the residual rate on all links. It gives the probability of success that there are r units of remaining bandwidth in link (k, l).

$$p_{k,l}(r) = e^{-\mu_{k,l}r} \tag{5.3.11}$$

where $p_{k,l}(r)$ is the distribution of residual rate r on all links, and $\mu_{k,l}$ is a constant value for link (k, l).

As we are interested in having an end-to-end QoS guarantee, from source s to destination t, then a traffic demand for a connection i from s_i to t_i has an upper bound for the delay as proposed by Guerin and Orda [9]

$$t_{upperbound_{c,i}}(P, r) = \frac{\sigma_i + n(P)v_{i,\max}}{r_{\min}} + \sum_{(k,l)\in P} t_{prop} \tag{5.3.12}$$

where σ_i is the bias value, related to connection's i maximal burst, $n(P)$ is the number of hops in path P, $v_{i,max}$ is the maximal packet size in connection i, and r_{min} is the minimal rate that can be guaranteed to the flow on each link along the path.

While the lower bound will be

$$t_{lowerbound_{c,i}}(P) = t_{lowerbound_{c,i}}(P, r(P)) = \frac{\sigma_i + n(P)v_{i,\max}}{r_{\min}(P)} + \sum_{(k,l)\in P} t_{prop} \tag{5.3.13}$$

This becomes the minimal guaranteed delay. A path in a connection i will only be feasible if the lower bound of time delay for a class c traffic in path P is

$$t_{lowerbound_{c,i}}(P) \leq t_{upperbound_{c,i}}(P, r) \tag{5.3.14}$$

and the minimum remaining bandwidth in path P is bigger than or equal to the bandwidth required by traffic class c (for link i in path P).

$$r_{\min}(P) \geq b_{requirement,c,i} \tag{5.3.15}$$

In our problem formulation (see (5.3.9)), we try to minimize the cost in z_{prop} and $z_{proc.}$. The first term, z_{prop} consists of the contribution from the number of hops $n(P)$, while the second term, $z_{proc.}$ concerns more to bandwidth perspective, $r(P)$. Both of these terms construct overall end-to end cost function from connection i. Therefore, the optimization criteria in (5.3.9) is to find a path for every incoming demand, which will minimize delay due to the number of hops in a path; and will maximize the residual bandwidth over the whole network which is related to transmission rate as in (5.3.10). The problem formulation then becomes:

$$\min(\sum_{k=1}^{n}\sum_{l=1}^{n}(\sum_{i=1}^{I}(t_{uplink,k} + t_{downlink,l} + t(P_{k,l})) + \frac{\sigma_{i,k,l} + n(P_{k,l})v_{i,\max}}{b_{k,l}})\nu_{i,kl}) \tag{5.3.16}$$

This strategy attempts to find a path that minimizes the propagation delay by looking for the minimal number of hops. It also tries to maximize the residual bandwidth.

As well as the above criterion, traffic allocation necessitates satisfying the following constraints:

Bandwidth constraint:

The capacity of the transmission link between satellites is limited. Therefore, total traffic in a particular link (k,l),v_{kl} should be lower than available bandwidth in this transmission link b_{kl}. Total traffic in this link is a summation of all connections i from source s to destination t of those who use the link (k,l), $i_{s,t}$

$$\nu_{kl} \leq b_{(k,l)} \tag{5.3.17}$$

$$\nu_{kl} = \sum_{c} D_{c,kl} \tag{5.3.18}$$

This means that $\nu_{k,l}$ (the total traffic load in the link (k,l) belongs to $D_{c,kl}$ (traffic demand for all classes of traffic c in the link (k,l)).

Node degree constraint:

Maximum flows that can be handled for each satellite define node degree constraint.

$$\frac{1}{2}[\sum_{k=1}^{N}(D_{c,kx} + D_{c,xk}) + \sum_{y \neq x} b_{xy}] \leq F_{\max} \tag{5.3.19}$$

where, $F_{\max}$ is maximum flows in one satellite. $D_{c,kx}$ and $D_{c,xk}$ give the traffic demand which is started in satellite k and ends in satellite x and started in satellite x and ends in satellite k respectively. The quantity b_{xy} is the remaining traffic demand which uses an ISL between satellite k and satellite x as the intermediate satellite (the source and destination is neither in satellite k or satellite x).

Data loss on path constraint:

In our model, a drop tail queue model is used, which drops packets if there is no available channel to allocate this packet. This model introduces a data loss into the model. Ansari, Arulambalam et al. in [132] model a queuing analysis, which is derived from the birth-death process of an $M/M/m/m$ queue model. In order to reduce the complexity of the problem, we consider using the drop tail queue model. Only data loss is considered, which corresponds to the condition where there is no available channel to satisfy a demand. Collision, which occurs when two packets use the same resource, is not considered in our model. Average blocking probability is limited by permitted blocking probability, which reflects performance of the satellite model.

$$\bar{p}_{loss(a,b)} \leq (1 - p_{\Delta loss}) \tag{5.3.20}$$

where

$\bar{p}_{loss(a,b)}$ = average blocking probability

$p_{\Delta loss}$ = permitted blocking probability

As a LEO satellite network has a dynamic topology, it introduces a complexity into the above criterion. The problem of handover becomes a significant issue in the LEO satellite network in order to ensure that the ongoing connection can maintain communication, and that the new arriving demand can be persuaded. Our proposed methodology to cope with these issues is discussed in the next section.

5.4 Methodology

LEO satellites move periodically in their orbits. Orbit periods of these satellites depend on their orbital position. Chang, Kim, et al. denote orbital period of LEO satellites as o_{LEO} and the earth orbital period as o_{earth} ; and express the system period as in [67]:

$$o_{system} = o_{LEO} \times o_{earth} \tag{5.4.1}$$

If we divide the system period into sufficiently small time intervals, so that we can make an assumption that an MT can use the same satellite for the whole time interval the MT does not need to be handed over to another satellite within this small time interval. We can assume that the satellite position is unchanged within this small time interval.

Since the satellite's position remains unchanged, it can be assumed that the satellite constellation topology is unchanged within this time interval. Changes in the satellite's topology occur only in transition of time intervals.. In this transition of time intervals, we locate satellites according to their new position in their topology and try to allocate the ongoing connection (traffic load) into this new satellite position. The satellite's position will remain unchanged until the end of the time period. In the next time interval's transition, the same procedure will be undertaken. Figure 5.4.1 as given in [67] shows this procedure in detail. We consider those satellites, only partially visible within a given time interval, as invisible. The intention is that an assumption can be made that the LEO satellite network has a fixed topology throughout a time interval. A routing algorithm can be performed in the LEO satellite network, in the same way as a set of routing algorithm in a fixed topology network.

Routing problems are separated into two phases. The first phase of routing problem deals with traffic allocation when the satellite's position changes. This phase takes place at the beginning of each interval. The second phase of the routing problem deals with traffic allocation where the satellite's position can be assumed unchanged. This phase takes place inside the time interval.

It is necessary to define the length of time intervals where we can assume that the satellite's position is unchanged. Time intervals are determined by the satellite visibility period, which is called the sliding windows of our satellite system. The length of the satellite sliding windows

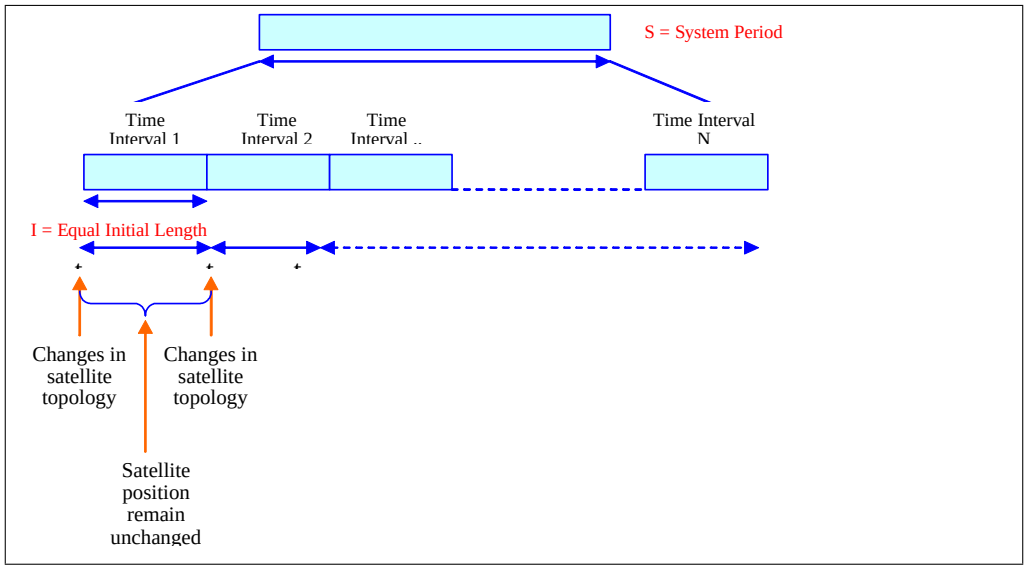

Figure 5.4.1: Periodical time division in equal initial length periodic

is dynamic, depending on the traffic load on the network.

5.4.1 Updating Sliding Windows

The length of the sliding window's period in our satellite system is updated to adhere to the average traffic load in the network. The length of the sliding window itself is only related to the orbital period of the LEO satellites and the LEO satellite altitude, and we call it the maximum length of the sliding windows. It defines the maximum length of time that a LEO satellite is visible to a user on the ground. In addition, we divided the maximum length of the sliding windows into smaller intervals, which are called the sliding window's length period. This sliding window's length period has a variable length, which depends on the average loaded traffic. The variable length has its maximum, which is called the sliding windows maximum. This will be explained further in the next paragraphs. We use Auto Regressive Moving Average (ARMA) with order 2 to calculate the new sliding window's length.

We assume that the traffic load Q_t is dependent on the traffic load of the last two time intervals Q_{t-1} and Q_{t-2} is.

The Auto regression of order 2 $AR(2)$ is:

$$Q_t = \mu + \alpha_1(Q_{t-1} - \mu) + \alpha_2(Q_{t-2} - \mu) + e_t \tag{5.4.2}$$

, The Moving Average of order 2 $MA(2)$ is:

$$X_t = \mu + e_t + \beta_1 e_{t-1} + \beta_2 e_{t-2} \tag{5.4.3}$$

So that the $ARMA(2)$ is:

$$Q_t = \mu + \alpha_1(Q_{t-1} - \mu) + \alpha_2(Q_{t-2} - \mu) + e_t + \beta_1 e_{t-1} + \beta_2 e_{t-2} \tag{5.4.4}$$

where α_1, α_2, β_1, β_2 are constant parameters, and μ is the mean as given below

$$\mu t = E(Q_t) \equiv \mu \tag{5.4.5}$$

and e_t is given as:

$$e_t = \sigma_t^2 = E^2(Q_t - \mu) \tag{5.4.6}$$

The number of time intervals inside the system period N_{new} is :

$$N_{new} = \left(1 + \frac{Q_t - Q_{t-1}}{Q_{t-1}}\right) N_{old} \tag{5.4.7}$$

with the new sliding window interval I_{New} as follow

$$I_{new} = \frac{o_S}{N_{new}} \tag{5.4.8}$$

where,

I_{new}, I_{old} = new and old initial length of sliding window

N_{new}, N_{old} = new and old numbers of time intervals inside system period

5.4.2 Satellite Allocation at The Beginning of Each Sliding Window

In the earth fixed cell based satellite model, the satellite's service coverage areas are occasionally overlapping. A MT chooses a satellite, which has enhanced signal strength. Due to the orbital movement of satellites, MTs retain a satellite in sight only for a few minutes, depending on the altitude of the satellite and its orbital period. Cruickshank, H., et al. in their BISANTE project

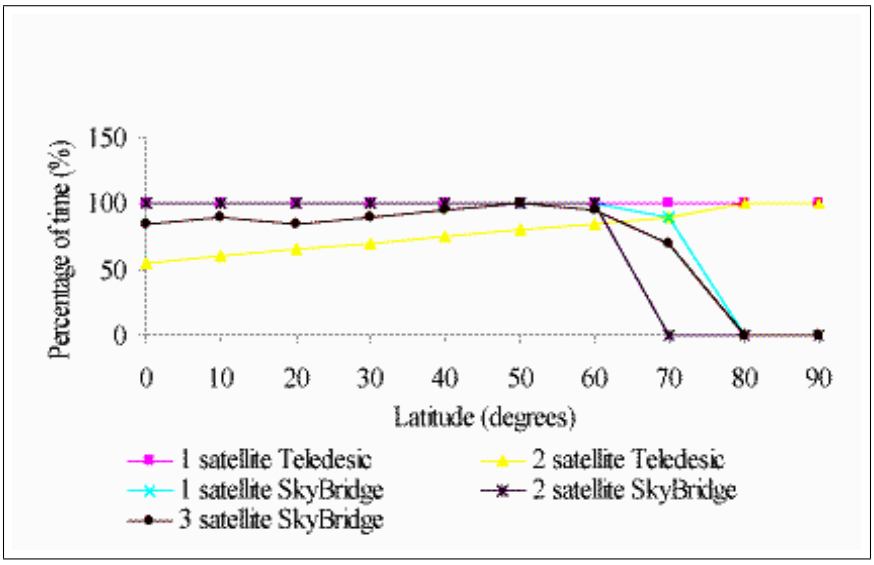

Figure 5.4.2: Satellite visibility of Teledesic and Skybridge from [54]

report [52] calculated the visibility of Teledesic and SkyBridge satellites from different locations on earth. Figure 5.4.2 shows the visibility time of Teledesic and Skybridge satellites from different locations. For other satellite systems, to calculate the visibility of different satellite constellations, the two line element set data is available at http://celestrak.com/NORAD/elements/. Figure 5.4.2 shows that every MT on earth from latitude ϕ=0^0 (in equatorial) to ϕ=90^0 (in polar region) can have 100% visibility of one Teledesic satellite, while, every MT on earth can only maintain 100 % visibility for one Skybridge satellite between latitude ϕ=0^0 to latitude ϕ=60^0. The visibility will diminish when MTs approach the polar region. 100% visibility of two Teledesic satellites is available when MT is in about ϕ=60^0 of latitude. Presume that the percentage of visibility of n satellites, from a MT with ϕ latitude, is $\delta_{n,\phi}$. In figure 5.4.3 A MT with latitude ϕ is at X on earth surface with O as earth centre. Assume that a satellite

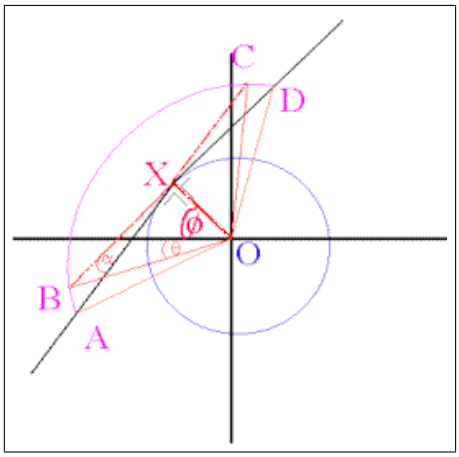

Figure 5.4.3: Visibility time interval of a satellite

moves in its orbit from time t_0 (at A, which is the beginning of horizon from the position of MT) to time t_4 (at D, which is the other end of horizon from the position of MT). Suppose that the satellite becomes visible to MT at time t_1 (at position B) and will disappear after reaching position C (at time t_3); the satellite becomes visible at time t_1, where the satellite has elevation θ and angle α $(\angle BXA)$ to the MT.

We are interested in the length of trajectory from t_1 to t_3, which is the time that a satellite is visible from MT. The related angle is $2 \times (\theta+\phi)$.

$$\frac{OX}{\sin(\angle XBO)} = \frac{OB}{\sin(\angle BXO)} \tag{5.4.9}$$

where, OX is the earth radius (R_{earth}) and OB is the satellite distance from earth centre $O(R_{earth}+R_{satellite})$. $R_{satellite}$ is the satellite altitude from earth surface. $\angle XBO$ is the total angle of $(\pi/2 - (\alpha+\phi+\theta))$

$$\phi + \theta = \frac{\pi}{2} - \alpha - \arcsin(\frac{R_{earth}\sin(\frac{\pi}{2}+\alpha)}{R_{earth}+R_{satellite}}) \tag{5.4.10}$$

The length of trajectory that satellite will be visible is then

$$2(\phi + \theta) = \pi - 2(\alpha + \arcsin(\frac{R_{earth}\sin(\frac{\pi}{2}+\alpha)}{R_{earth}+R_{satellite}})) \tag{5.4.11}$$

The maximum sliding window (in minutes) then becomes

$$Slidingwindow_{\max} = (\pi - 2(\alpha + \arcsin(\frac{R_{earth}\sin(\frac{\pi}{2}+\alpha)}{R_{earth}+R_{satellite}}))) \times \frac{o_{LEO}}{2\pi} \times \delta_{n,\psi} \qquad (5.4.12)$$

in which the satellite moves with the angular velocity of *(2π)/ O_{LEO} rad/minute.*

For example for an IRIDIUM constellation O_{LEO} =102 minutes, $R_{satellite}$ = 780 Km and the R_{earth}= 6376 Km.

The satellite is invisible for a fraction of time, which is the area of the sphere defined by the satellite orbits:

$$T_{invisible} = \frac{4(R_{earth}+R_{satellite})}{4\pi(R_{earth}+R_{satellite})^2} \int_0^1 \int_0^1 \frac{dxdy}{sqrt((R_{earth}+R_{satellite})^2 - x^2 - y^2)} \qquad (5.4.13)$$

Supposing that the satellite will immediately become visible on the horizon and disappear at the other end of the horizon (α=0), presume that the visibility of one (n=1) satellite from every latitude on earth surface is 100%. The maximum sliding window should be about 15 minutes.

In figure 5.4.4, there are five different altitudes of LEO satellites: 300km, 600km, 900km, 1200km, and 1500km. The orbital periods of these LEO satellites are given respectively: 90minutes, 96 minutes, 102minutes, 109 minutes, and 115 minutes. In the case that the LEO satellite with altitude of 600km has a full visibility (100% visibility), then the maximum sliding window of this type of satellite is 14 minutes. In case that the LEO satellite has only 75% visibility, then the same satellite will have only 11 minutes of its sliding window.

Sliding window length suggests that we should apply an updating period smaller than this maximum sliding window, to ensure that the ongoing connection can be maintained when the satellite becomes invisible.

Prior to the disappearance of this satellite a new connection needs to be established. A handover procedure to the new approaching satellite needs to be constructed. In case we only have one visible satellite at a time, the handover procedure needs to be initiated before the currently used satellite becomes invisible. This case is shown in figure 5.4.5. A connection between $MT1$ and $MT2$ is constructed using satellite 1($Sat1$) and satellite 2($Sat2$), via ISLs

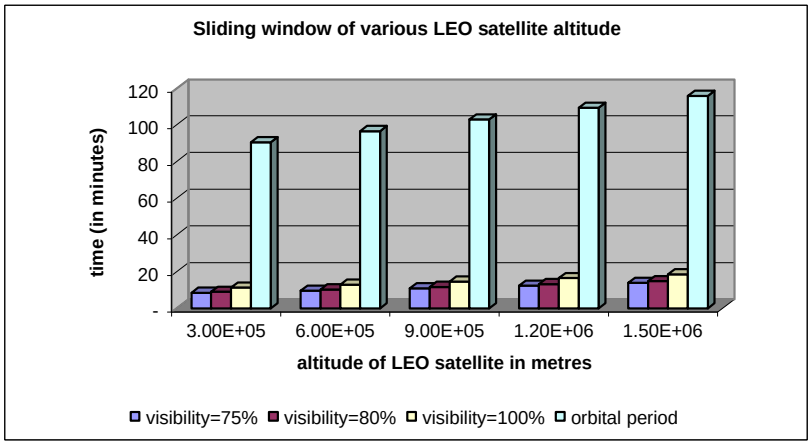

Figure 5.4.4: Maximum sliding window of various LEO satellite altitude with various percentage of visibility

between these two satellites. $MT1$ is currently inside the service coverage of $Sat1$, while $MT2$ is located inside $Sat2$'s service coverage. Both satellites move following clockwise direction as given in figure 5.4.5. $MT2$ starts losing service coverage from $Sat2$ and starts to obtain $Sat1$'s coverage. Since currently $MT2$ is only covered by $Sat2$, then it needs to start the handover procedure to allocate the connection to the next available satellite, $Sat1$.

In the case that two or more satellites are visible at the same time, a soft handover can be used to allocate the ongoing call into another visible satellite, which has a better signal performance or better visibility than the current one. This case is shown in figure 5.4.6. $MT2$ is currently in the coverage of $Sat3$ and $Sat4$. Soon, $Sat3$ and $Sat2$ will cover $MT2$. Every time there should be two visible satellites for each user. $MT2$ compares signal strength between these two satellites and chooses the best satellite to build a connection.

5.4.3 Handover

Satellite links in the LEO satellite constellation are constructed from two different types of ISLs. The first one is the intraplane satellite connection. This is an inter satellite link which is constructed between two satellites in the same plane; whereas, interplane satellite connection is an ISL which is constructed between two neighboring satellites in a different plane.

The intraplane satellite connection remains continuous for the whole period of time, while interplane satellite connection in some LEO topology will be turned off for a period of time.

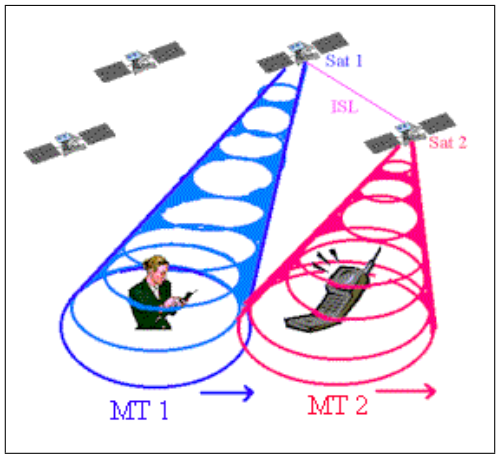

Figure 5.4.5: Handover procedure between neighboring satellites

For example, in the near polar constellation, intraplane satellite connection is switched off in the polar region, to reduce the signal interference. In the case of IRIDIUM intraplane ISLs will only be maintained between latitudes of approximately 60° north or south of the equator [58].

A first degree ISL only provides a connection between a satellite with its direct neighbors. A second degree ISL will provide a connection between a satellite with not only its direct neighbor satellites, but also with the next neighbor satellites as shown in the following figures (figure 5.4.7). $Sat1$ in the first degree ISL has 4 neighboring satellites, which $Sat1$ can directly communicate with. In the second degree ISL, $Sat1$ has more neighboring satellites, with which $Sat1$ can communicate (figure 5.4.8). ISL connections will be turned off whenever the satellites are in the polar region or when the satellites are in the seam. Satellites will be in the seam, if the satellites are in the neighboring planes and the moving direction of the satellites in a plane is in counter rotating with the satellites in the neighboring plane.In figure 5.4.9, the satellites' positions are given in their orbital plane. There are 6 planes with 11 satellites on each plane. Satellites in the third plane move in a counter rotating direction to satellites in the fourth plane. Satellites on these planes are in the seam with each other. Hence, in some LEO satellite systems, ISLs between those satellites on these neighboring planes are turned off.

In figure 5.4.10, ISLs between neighboring planes are given. ISLs between satellites on the third plane and satellites on the fourth plane, and ISLs on the polar region are usually turned off. Besides intersatellite handover, there is another type of handover, which is called

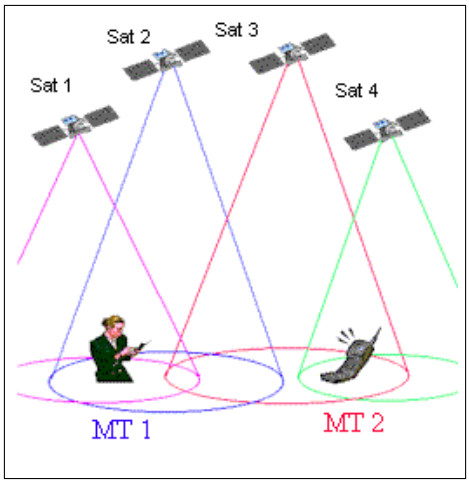

Figure 5.4.6: Soft handover procedure

intrasatellite handover. This is a handover between spot beams in one satellite. The last type of handover is a handover that occurs once the ISLs between satellites (either on the seam region or near polar region) are turned off. In our model, we only consider the intersatellite handover. The other two types of handover will be the interest of our future research.

Akyildiz, Uzunalioglu et al. discussed handover procedures in their paper [58, 62]. They propose Footprint Handover Rerouting Protocol (FHRP), which has two phases: route augmentation, and rerouting. FHRP has a goal to find an optimum route without performing any finding algorithm after a handover. A handover is necessary as soon as one of the end satellites, either the source or the destination satellite, is no longer visible or available for the MTs, or because the ISL links are turned off [58, 62].

We apply an almost comparable handover procedure, except that in our model, we consider handover of the ongoing connection to the next visible satellite on the same plane. This will reduce the complexity when the satellites are in the polar region, or when the satellites are in the seam orbital planes. Figure 5.4.11 shows a case in which an ongoing connection from source s (*satellite2*) to destination t (*satellite4*) needs to be handed over. Due to their movement, satellites become invisible from MTs past the maximum sliding window time. A handover needs to be performed to maintain this connection. Since we know the satellite's orbital direction (as given in figure 5.4.11), the next visible satellite for s will be *satellite1*, and the next visible

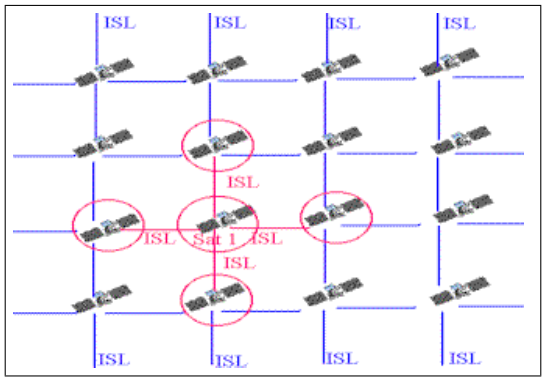

Figure 5.4.7: One degree ISL

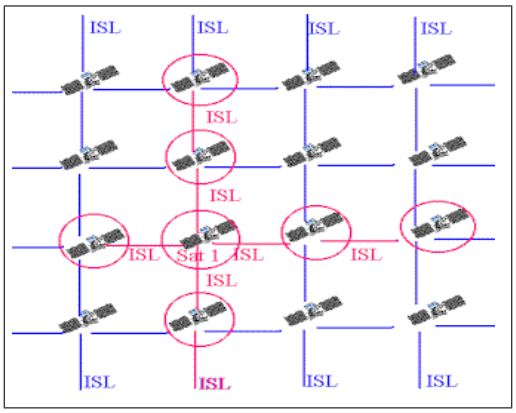

Figure 5.4.8: Two degree ISL

satellite for t is *satellite7*. Both satellites, *satellite1* and *satellite7* are on the same orbital plane with *satellite2* and *satellite4*, respectively. Intraplane handover will be performed from *satellite2* to *satellite1* and from *satellite4* to *satellite7*. The resulting new connection after the handover is shown below:

$$P_{new} = l_{S_{new}S_{old}} + P_{old} + l_{t_{old}t_{new}} \tag{5.4.14}$$

The new connection P_{new} will be constructed from firstly the old connection P_{old}; secondly by adding the ISL between the new source and the old source $(l_{s_{new},s_{old}})$; and finally by adding the ISL between the new destination and the old destination $(l_{t_{old},t_{new}})$. In our example, the new connection will be $1, 2, 3, 4, 7$.

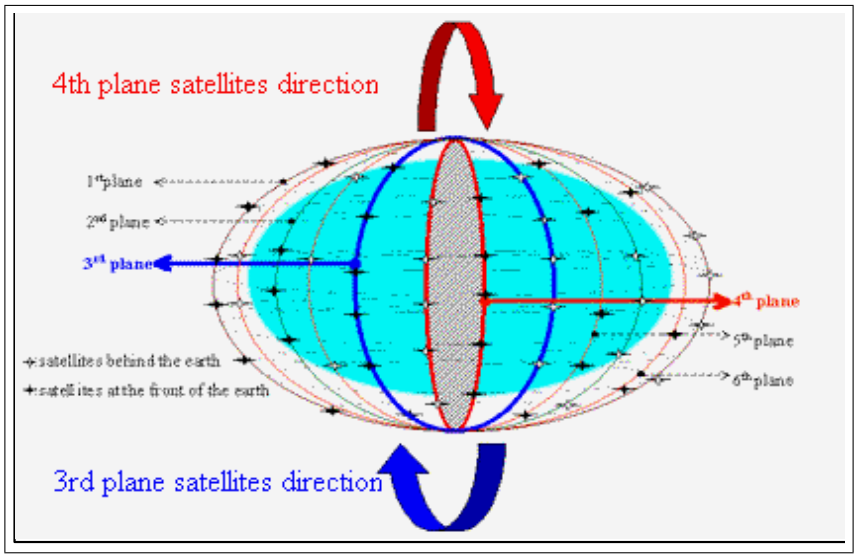

Figure 5.4.9: Satellites constellation with their orbital position and direction

5.4.4 Different Types of Traffic Classes

Since we are dealing with the demand of various traffic classes, we introduce a mechanism in which more privilege will be given for higher priority traffic. In order to do this, we introduce a parameter λ^c into our problem formulation. λ^c is a privilege parameter for traffic class c, which will set up the objective function (5.3.9) of our routing allocation problem as shown below:

$$\min(\sum_{k=1}^{N}\sum_{l=1}^{N}(\sum_{c=1}^{C}\sum_{i_c=1}^{I}(z_{prop_{kl}} + \lambda^c \ z_{proc_{kl}})v_{i,kl})) \tag{5.4.15}$$

With class c, $0<c \leq C$ in a satellite network which supports C classes of traffic. In the second term, in (5.4.15), we introduce a notional cost based on the unused capacity of the transmission link. As available bandwidth of a transmission link is limited, it is important to share traffic over the whole network. When all links have a spare capacity, then the ability to handle additional demand is enhanced. Hence, when adding a new demand to an existing pattern of traffic, we assign a higher cost if the traffic is allocated to a link with little remaining capacity. This tends to divert new traffic to paths with spare capacity. The two terms in objective function may sometimes conflict, since the minimum hop route will not necessarily minimize the cost due to

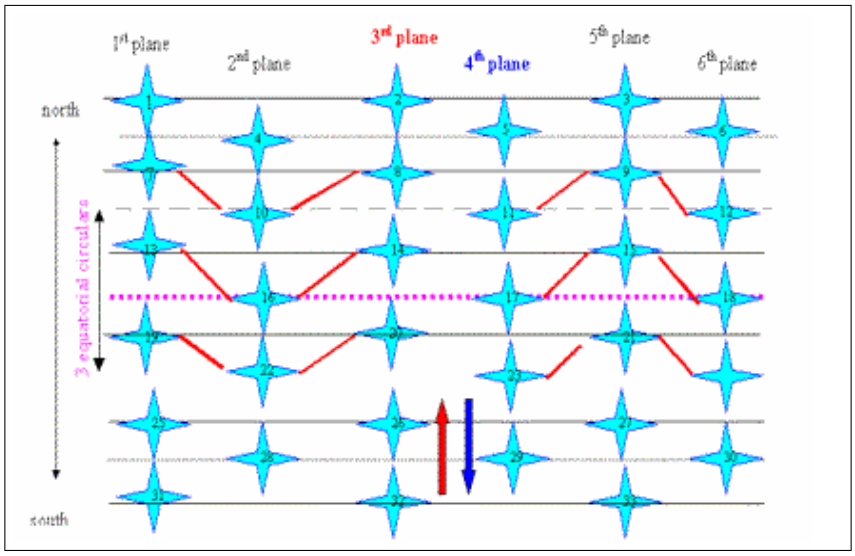

Figure 5.4.10: ISLs on the seam region are turned off

the bandwidth available. By introducing λ^c in (5.4.15) we would like to give a preference to a high priority class of traffic and provide a mechanism to give preference to lightly-loaded links and allow a better control of traffic distribution. We use a large value of λ^c for a low priority class of traffic, and a small value of λ^c for a high priority class of traffic. By assigning a large value of λ^c to a low priority class of traffic the low priority of traffic is diverted to lightly-loaded ISLs. Hence, we can reserve some amount of bandwidth for high priority traffic. The choice of the value λ^c depends on how much reserve space should be allocated to the high priority class, and also, on the maximum hops allowed to re-route the low priority traffic. In our research, we consider 2 classes of traffic, namely the high priority traffic and the low priority traffic.

5.4.5 Analysis of Implementation

Handover and channel assignment problems should be considered simultaneously, in order to obtain an optimal policy, which takes into account cost function, blocking rate and call quality in the satellite system.

Due to the relative movement of satellites with respect to mobile users, several satellite handovers are necessary during a voice call. Since satellites are traveling along their orbit, connection of any user to the satellite must be handed over to a new satellite footprint, although the

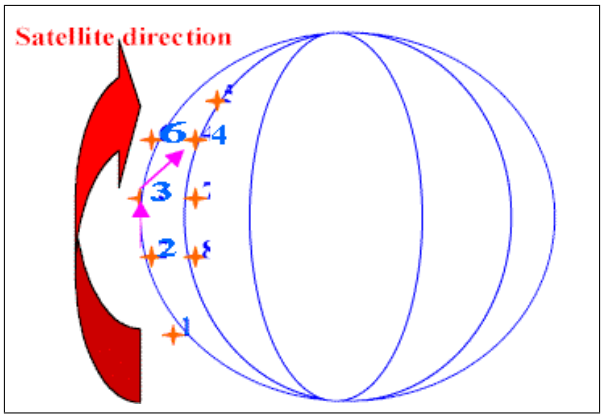

Figure 5.4.11: Intra plane handover procedure

user has never moved during the call session. The handover request has higher priority over new arrivals. Traditional handover schemes for terrestrial cellular networks are based on threshold policies. Recently, handover phenomenon was formulated as a reward/cost optimization problem. The received signal is considered as a stochastic process with an associated reward, while the handover is associated with a switching penalty.

The message of each active user is divided into packets with identical length, and time is divided into slots. The duration of these slots is equal to the transmission time of one packet. Signal strength between MT and the satellite is measured periodically at the beginning of the time interval, in order to make a handover decision.

The handover decision and channel assignments are made after each measurement is made. In our simulation, we use the distance between MT and the satellite as the basis of handover decisions. The handover is initiated as soon as another satellite has a closer distance to the MT than the current satellite. Satellite candidates must have a minimum elevation, θ_{min} to establish a link with this MT. Links between MT and a satellite, which have low elevations are easily shadowed. It is assumed that the variation of satellite elevation only comes from satellite movements. Movements of MT have no effect on the elevation angle.

We follow the UMTS handover procedure, which consist of the following three phases. The first phase is the information gathering phase (handover initiation phase). In this phase, information about signal strengths and location of satellites is collected. The second phase is

the decision phase. Based on the gathered information, decisions will be made regarding which connection needs to be handed over, and which satellite has the ability to preserve the QoS guarantee. The last phase is the execution phase. In this phase, the actual handover procedure is performed.

5.5 Summary

In this chapter, we defined the optimisation problem, which occurs when we allocate different types of traffic classes into a dynamic topology of LEO satellite networks. Firstly, we considered the characteristics of dynamic LEO topology with their satellites are moving in a predictable way. We used this information to predict the satellite position in the next time interval. We divided the orbital period of satellites into small time intervals. We made an assumption that the satellite position will be unchanged inside these time intervals. The satellite position changed only at the beginning of each time interval. Our objective function consisted of two parts. The first part is due to the weighted delay in propagation time, and the second part is due to the remaining bandwidth. In order to reduce the complexity in the case of the handover, we considered the handover of an ongoing connection to the next visible satellite in the same plane. In addition, we keep the middle part of the ISL's path between the source and the destination. We added an extra new path between an old source to a new source and an extra new path between an old destination and a new destination, respectively. Moreover, since we are dealing with a multimedia type of traffic, which consists of two types of traffic (high and low priority traffic), we introduced a privilege parameter. This parameter is used to give more privilege to the high priority traffic and to distribute the traffic more evenly over the satellite network.

Chapter 6

ALGORITHMS

6.1 Introduction

In this chapter, we discuss various algorithms which are related to our traffic allocation problem in the LEO satellite. Much research has been undertaken into traffic allocation algorithms, the objective of which is to allocate traffic demand in different networks. In the following section, an overview of various traffic allocation and scheduling algorithms that have been studied is given. Thereafter, an overview of basic algorithms which are used to perform our traffic allocation algorithm is given; namely Genetic Algorithm (GA), Linear Programming (LP), Tabu Search (TS) and Dijkstra Shortest Path algorithm (SP).

6.2 Various Traffic Allocation Algorithms

There is a change of characterization in communication traffic from traditional traffic, which is dominated by voice service, into a multimedia traffic which has video, voice, and other data traffic. This change the way that traffic should be allocated in a communication network, from a best effort based traffic allocation to a QoS based traffic allocation. This, in turn, changes the way that a routing algorithm is implemented in the communication network. Lee Breslau and Shenker compared the two different type of traffic allocation, best effort service and reservation-capable service to guarantee a QoS [33]. Paschalidis and Tsitsiklis propose a routing algorithm which has a congestion dependent cost as a way to provide QoS guarantee. They considered

an exact and approximation approach of computation. They demonstrated that by using an approximation approach a reasonable performance of routing algorithm could be obtained [34].

Bremler-Barr, Afek et al. illustrated another type of algorithm which schedules traffic from node to node. They introduce a clue into the *IP* header. In this approach, they add an extra 5 bits in the IP header to tell its downstream router where a good point to start for the IP lookup is [40]. Lai and Chang, Orda, Shaikh, Rexford et al. studied other types of IP based routing algorithm in ATM networks environment [41–43]. While Chlamtac and Farago, Kwon, Choi et al., Lu, Bharghavan et al., Sarikaya et al. analyzed routing algorithms in a wireless environment. If a routing algorithm is used in a wireless environment, unique characteristics of wireless media ('bursty' channel errors and location dependent channel capacity and errors) should be introduced [4,44–46]. Various adaptive routing algorithms perform a fair scheduling of delay and rate-sensitive packet. In [133], the authors propose a scheme of predictive bandwidth allocation strategy that exploits the topology of the network and maintaining high bandwidth utilization. The results showed a low call dropping probability, and provide a reliable handoff of ongoing calls. The authors divided the reservation scheme into two strategy, fixed reservation and predictive reservation. In the fixed reservation, available bandwidth is permanently reserved for handoff. While in predictive reservation, available bandwidth is reserved using a probabilistic approach. Through adaptive routing algorithm, traffic allocation in a network could be updated according to the current condition of the network. These adaptive routing algorithms are studied in [6, 47–49].

In the satellite network traffic allocation problem, some studies have been done to find the best way of allocating traffic demand into satellite links. Throughput evaluation and channel assignment for a Satellite Switched CDMA is investigated in [134], in which the uplink and downlink channel assignment is considered in terms of space, frequency and code division. Traffic allocation for the LEO satellite system is investigated in [61], in which different types of queuing systems are used such as First Input First Output (FIFO), Last Useful Instant (LUI) - based on the maximum time that a handover needs to be accomplished. In this paper, research is focused on the handover performance and the queuing process before assigning a channel. Fixed Traffic allocation with Queuing Handover (FCA-QH) requests and Dynamic Traffic allocation with Queuing Handover (DCA-QH) requests are discussed. These show that

DCA obtain significant improvement in terms of maximum traffic intensity per cell and capacity per cell. A performance study of the LEO satellites in terms of system capacity and average number of handovers is studied in Ganz, Gong et al. paper [28]. This study focuses on the IRIDIUM system. Policies for handovers and channel assignment in the LEO satellites, by considering the LEO satellites as bent-pipe transponders are investigated in [135]. A finite-horizon Markov decision process is formulated using the probabilistic properties of signals and of the traffic in the footprints, which have an objective to minimize the switching costs and the blocking costs of traffic. Traffic management in a GEO satellite consisting of ground stations, which perform a mesh connected topology is studied in [132]. The optimization used two neural network based optimization techniques: simulated annealing and mean field annealing (MFA), which show a better performance than the pure dynamic routing with a fixed configuration as used by AT&T's Dynamic Non hierarchical Routing (DNHR) method, and Canadian Telecomm Dynamic Control Routing (DCR).

In a satellite communication, one of the most significant parameters of QoS is the delay time. Therefore, delay time becomes an important factor in a routing algorithm in a satellite communication network. There has been some research undertaken considering a routing allocation problem and admission control in satellite communication. In [136], the authors derived a new algorithm called Gauge & Gate Reservation with Independent Probing (GRIP), which proposed a solution for the admission control problem in a heterogeneous network, comprised of satellite and terrestrial network connected to an IP core network. GRIP is intended to operate over Diffserv Internet, and is composed of three components: GRIP source node protocol, GRIP destination protocol, and GRIP Internal router decision criterion. In a yet to be published paper [137], the authors develop a routing and scheduling algorithm for packet transmissions in a LEO satellite network. In this paper, they consider three transmission scheduling schemes to decide which packet they will route first: the random packet win, oldest packet win, and shortest hop win. The winning packet will have the highest priority to be routed to its destination. The routing algorithm, which is used in both papers, is the shortest path algorithm. Ekici, Akyildiz et al. consider a Border Gateway Protocol Satellite version (BGP-S). In BGP-S a new path from source to destination is discovered, by measuring the delay from BGP-S to the destination. This delay information remains local to the BGP-S protocol. If alternative paths are available,

the choice is based on the delays on the existing paths [76]. Sun and Modiano evaluate different aspects of primary capacity and spare capacity for recovering from a link or node failure. This is useful in case of handover in the LEO satellite network. In general, in case of a link failure, a restoration scheme can be classified as link-based restoration whereby affected traffic is rerouted over a set of replacement paths through the spare capacity of a network between the two nodes terminating the failed link or as path-based restoration whereby affected traffic is rerouted over a set of replacement paths between their source and destination [138].

The link based scheme is significantly simpler and faster to recover than the path-based scheme. On the other hand, the amount of spare capacity needed for the link-based scheme is greater than that of path-based restoration, since the latter has the freedom to reroute the complete source to destination path using most efficient backup path. Some aspects mentioned above such as adaptive, link based and path based restoration, the use of alternative paths available in a local table (such as BGP-S) is accommodated into our routing algorithm. Our combination algorithm, GALPEDA, applies these properties to enhance the routing performance. Before we discuss different algorithms, which are used as the base of our GALPEDA algorithm, a brief explanation of our combinatorial optimization problem is given.

The problem of finding an optimal traffic allocation in a satellite network, while respecting some constraints, can be considered as the combinatorial optimization problem. Suppose that there is a computational problem $g(n)$, which has an input domain set ι_g of instances and for each $x \in \iota_g$ there is a corresponding solution $answer_g(x)$. A feasible solution for the problem $g(n)$ is a subset $\upsilon_g(x) \in answer_g(x)$.

According to Corne, Dorigo et al., an algorithm attempts to solve $g(n)$ of an input $x \in \iota_g$ by finding an output $y \in \upsilon_g(x)$. If $\upsilon_g(x)$ is empty an algorithm should be able to tell us that there is no such output $y \in \upsilon_g(x)$. Combinatorial optimization problem is a special kind of search problem where every instance $x \in \iota_g$ has a set solution $\upsilon_g(x)$, which satisfies an objective function and a goal [139].

Aarts and Lenstra define a combinatorial optimization problem as: 'A combinatorial optimization problem is a problem of decision making in case of discrete alternatives and solving them to find an optimal solution among a finite numbers of alternatives'. A combinatorial optimization problem is specified by a set of problem instances and it can be defined as a mini-

mization problem or a maximization problem. A decision needs to be made based on the sum of costs criterion, which will provide a quantitative measure of the quality of each solution [140].

Optimization problems in satellite networks belong to this combinatorial optimization problem. As the complexity of the decision in satellite traffic allocation problems could not be solved by a deterministic Turing machine in polynomial time [141], this combinatorial optimization problem will belong to the class Nondeterministic Polynomial-complete (NP-Complete) problems. The design problems of choosing a set of links for a given set of nodes to satisfy some cost constraints is considered as NP-hard [142]. Therefore, the complexity of this optimization belongs to NP-hard.

According to Mitchell, there are three different approaches to solving this NP-hard type of problem. We can choose an enumerative method that will guarantee finding an optimal solution, but the processing will probably be impractical. A different approach is found by using an approximation algorithm that runs in polynomial time, which will attempt to locate the optimal solution in respect to some constraints. The last approach is by applying some type of heuristic technique, without any guarantee in terms of solution quality or running time. Metaheuristic will refer to a strategy to guide and modify other heuristics to produce a better solution than the solution, which is normally generated in a guest for local optimality [143]. Usually, in metaheuristic an adaptive memory, neighborhood exploration, and a method of carrying the current solution throughout one iteration to another will be used.

In this approximation approach, there are two classes of approximation algorithms according to Aarts and Lenstra: constructive and local search. In constructive algorithms, searching for an optimal solution starts with an empty space of solution. Iteratively searching procedure composes the solution. While the local search algorithm starts with an initial population, searching procedure explores a superior alternative solution in neighboring space of the initial population. Some algorithms, which belong to this local search class, introduce a distinctive characteristic into their searching procedure in which an alternative solution can be explored in the other neighboring space. This is an extension of local search algorithms, by performing the first approach: a multistart approach. In this approach, a simple local search algorithm is performed several times using different initial solutions as the starting point and keeping the best solution found as the final solution. The second approach is the multilevel approach. In

this approach, an iterated local search algorithm is performed, in which the starting point of subsequent local searches is obtained by modifying a local optima from the previous run. The last approach is a search strategic approach, which is a search strategy that is performed to find a cost-decreasing neighbor [140].

In the local search, neighborhood is a significant aspect. The local search algorithm begins with an initial solution and then iteratively attempts to find better solutions by searching the neighboring space. Three basic steps need to be performed: generation of an initial solution, generation of a neighboring solution, and cost calculation of the solution.

In Sedgewick, performance of combinatorial algorithms is distinguished in the worst case and the average case. In the first case of performance analysis, we need to ignore constant factors to be able to determine the functional dependence of running time on different types or number (size) of inputs, n. 'A function g(n) will have order of computational complexity of $O(f(n))$ if there exist constants x_0 and n_0 such that $g(n)$ is less than x_0 $f(n)$ for all $n>n_0$. The worst case running time is the maximum time that the program would need to execute for an input of size n'. Introducing this worst case performance analysis will allow us to define the upper bound of the computational complexity of an algorithm. In the second case, the average case, we calculate the average number of times each instruction is executed. The total time will be the summation of time required for performing all of these instructions together [144].

The performance analysis of an approximation algorithm can also be quantified by its running time and its solution quality. The running time will be given by the number of CPU seconds. The quality of the solution will be measured by the ratio of its cost value to that of an optimal solution or some bound for the optimal value.

Besides these performance analysis measurements, if we choose an algorithm, some additional considerations need to be looked at. The first consideration is to find the simplest algorithm, which can solve a given problem. A very careful investigation needs to be made for the bottlenecks in the system. Especially in large systems, it is often the design requirements of the system that dictate from the start which algorithm will be effective. Relatively faster algorithms are often more complicated, but sometimes it is not much more complicated than a slower one. Therefore, we need to consider whether we would like to deal with the added complexity to increase the speed of our algorithm. Another factor is the reliability/robustness,

which is a measurement of whether the algorithm will work correctly for different types of input data.

With these considerations in mind, we propose a combination algorithm to solve the traffic allocation problem in satellite network. We consider a combination of GA, LP and Extended Dijkstra shortest path Algorithm (EDA). We use GA to solve a global traffic allocation problem to improve the solution's quality. A mutation property of GA is used in order to escape from the local optima. LP is used to perform the crossover of GA. EDA is used in order to solve the local traffic allocation problem in a very short running time. In GA we use the first type of approximation approach. First we start with an empty space and iteratively we generate an initial population. In addition to that a multilevel approach is used between the time intervals, wherein the solution of previous time interval is modified and inserted into the new initial population. We consider a combination of these algorithms to find a near optimal solution within a reasonable running time. We propose in our research a combination algorithm to solve the traffic allocation problem in satellite networks in two ways.

The first traffic allocation problem will occur at the beginning of the time interval, when the satellite's positions are updated. And the second traffic allocation problem will occur inside the time interval, when we can assume that the satellite's position remains unchanged. We consider four different algorithms, which have useful properties in our satellite environment: the genetic algorithm (GA), the Linear Programming (LP), the tabu search (TS) and the Dijkstra shortest path algorithm (SP).

6.3 Genetic Algorithms

Some researchers realized that they could use the idea of real evolution in computer programming. According to Miller, in the mid 1960's, Lawrence Fogels suggested a new type of programming, evolutionary programming. They suggested a simulated evolution for simulated intelligence [145]. Later, according to Goldberg, John Holland and his students at the University of Michigan developed GA in the late 1960s and published in 1975 in his paper "Adaptation in Natural and Artificial Systems". The distinguished property of GA rather than other optimization and search procedure is that instead of working with the parameters themselves, GA

works with a coding of the parameter set. Moreover, GA starts the search from a population of points instead of only a single point. GA uses the objective function information and probabilistic transition rules instead of using deterministic rules [146]. In the 1970s, two different evolutionary algorithms, the GA of Holland and the evolution strategies of Rechenberg-Schewel merged [140]. The algorithm of Holland is more in adaptation, which emphasizes the importance of recombination in large populations, while Rechenberg-Shcwefel investigated mutation in very small populations for continuous parameter optimization. Recombination is more a global search based on restricted chance, while mutation is based predominantly on arbitrary chance. The GA mimics, as the first approach in evolutionary algorithms, processes in biological evolution with the ideas of natural selection and survival of the fittest and so provides effective solutions for an optimization problem. Deboeck et al. compared the GA with neural networks, he found that the GA can be more powerful than neural networks or other machine learning techniques [147]. Some research has been done by using GA for optimization in telecommunication networks. Chou et al [148] uses GAs for solving a network optimization problem by considering it as a degree-constrained minimum spanning tree problem.

6.3.1 Description

In a GA, we call a candidate solution a chromosome. A chromosome is encoded and assigned to a fitness value depending on the problem specific function. Binair representation of the chromosome is the most popular encoding method. A collection of chromosomes, which are candidate solutions for the problem, compose a population of solutions or chromosomes. This population will be maintained at each iteration step. At each iteration step the breeding pairs will be chosen from this population and a new child (solution) is created. The breeding process continues until a stopping criterion is reached.

There are three significant operators in GA. The first operation is selection. In this procedure the operator selects chromosomes in the population for reproduction. The selection depends on the fitness value of the chromosome: the fitter the chromosome (higher value) the more likely it will be selected to reproduce. The second operation is crossover. In this crossover, the operator performs according to some methods, depending on the encoding of the chromosomes, the actual breeding process. The last operation is mutation. In this procedure the operator

helps GA to escape from its local optima, and a modification is applied to individuals in the population.

6.3.2 Development

The GA is considered in our thesis as the basis of our combination algorithm, since the combination of GALP provides a global solution of the satellite network constellation problem. The solution of GALP is used to maintain the evenly distributed traffic allocation all over the globe. Introduction of our privilege parameter and moreover, the mutation in GA results in an algorithm which can escape from a local optima. This is significant in case of regional collision.

Previous work in [142, 149] used GA to optimize the design of communication network topologies, in which they used a collection of links. The initial population is generated using a heuristic developed for Minimum Cost Network Synthesis Problem (MCNSP) [149].In our

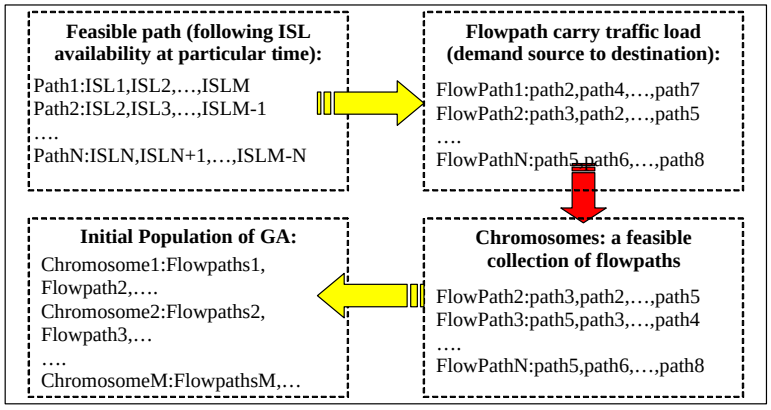

Figure 6.3.1: Initial population of Genetic Algorithm

research chromosomes will be represented as a collection of flow paths (see figure 6.3.1), with each path carrying a corresponding quantity of traffic from a source to destination. We assume that all links are bi-directional. In addition, since we are considering the LEO satellite communication network as the environment of this algorithm, the information about the satellite in latitude and longitude is recorded in each node data. This node data changes each time a particular time interval is passed by. The heuristic starts with an empty set of paths (this is the constructive part of our algorithm) and at each phase, we add the lowest cost path between

an origin and a destination, which carries traffic between these nodes.

We start with a permutation of all available satellites in the satellite network. We consider links between two satellites, which do not have an ISL as of having zero capacity. A chromosome consists of a permutation of these satellites numbers. Flow paths for traffic from source to destination at each satellite will be found by using a shortest path algorithm. We allocate the largest possible traffic that can be accommodated by this path, which will be defined by the traffic demand, the lowest unused capacity of the links and the lowest maximum degree of nodes in this path. If the maximum link capacity is reached, then this link becomes unavailable. If the flow capacity of a node is reached, we remove all of the links from this node. If we reach the maximum degree of a node, then unused links will be removed from that node. We terminate the heuristic when all of the traffic is allocated or there is no feasible solution, due to the infeasibility of the problem or due to the inability of the heuristic to find a feasible solution.

We perform a search to find a number of feasible solutions and use these solutions as the initial population. This initial population becomes a starting point for the GA. In addition, we provide in this heuristic our privilege parameter, which allocates the traffic load in relation to the class of traffic, in which a privilege for a class of high priority traffic is given. By introducing this privilege parameter in combination with characteristic of GA, a spare capacity is allocated for a high priority class of traffic. In addition to that the traffic is distributed more evenly in the satellite network. Since this GA is performed at the beginning of each time intervals, it considers the global traffic allocation problem. As given in the figure below, the traffic distribution is usually overloaded in the region between 60^0 and -60^0 of latitude (see figure 6.3.2). By introducing a GA and our privilege parameter, the traffic load is spread over the globe more evenly.In the previous work [151, 152], the selection procedure is performed by the random choice of two chromosomes from the initial population to become the parents. The two chromosomes are unified and delivered to the LP solver [152]. In addition to that there is no consideration regarding the dynamic properties of the network. Since in our research we must cope with a dynamic traffic allocation problem, we introduce new properties. At the beginning of a new time interval, we use a modified solution of the previous time interval as the first solution in the new initial population. Instead of starting with an empty space of population, we have one solution in the initial population. This first solution originates from

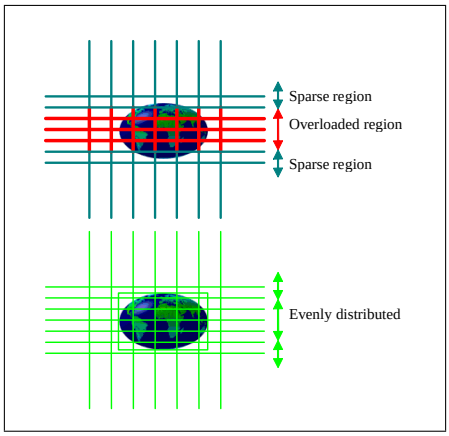

Figure 6.3.2: Traffic load in overloaded region and more evenly distributed traffic load

the previous time interval's solution, in which we replace the satellite positions in this previous time interval's solution with a new prediction of satellite positions. Introducing this solution, we are gaining historical information of traffic load in the previous time interval. Moreover, in the crossover procedure, we add a 'third' parent, which will give certain information regarding the previous time interval solution. This increases the speed so that the next LP processing time needs to find the best solution. With a small probability, we introduce a mutation property into the crossover procedure. The mutation is done by using a 'third' parent instead of a modified solution: the unmodified solution.The crossover will be performed inside the LP solver, and will deliver a new child as the result. The result is that the LP solution becomes a new candidate solution. If this new child has a lower cost or has a better fitness value than the worst solution in the initial population, then the new solution will replace the worst solution in this initial population.

6.4 Linear Programming

In 1990, a workshop called the Great Texas Traveling Salesman Problem (TSP) attended by participants with their own software to compete on TSP, resulted in the domination of Linear Programming (LP) based optimization techniques in the local search approaches. 'Linear programming copes with the problem of finding a vector x that minimizes a given linear function

$c^T x$ where x ranges over all vectors satisfying a given system $Ax \leq b$ of linear inequalities. We try to find a vector x which satisfies $Ax \leq b$ and has the smallest value of $c^T x$ ' [140].

In the late 1940s, G.B. Dantzig designed a fast solution method to solve the problem of LP called the simplex method. Simplex method is a method of making a trip along the vertices and edges of the polyhedron $\{x : Ax \leq b\}$, until an optimal vertex is attained. In 1979, L.G. Khachiyan solved LP-problem in polynomial time, but it was very slow. In 1984, N. Karmakar improved this speed [153].

6.4.1 Description

In our satellite network problem, the LP objective function is to minimize the total cost of allocating traffic demand into the ISLs, respecting some constraints. These constraints construct an Integer LP's (ILP) matrix (table 6.4.1) with Left Hand Side (LHS), Right Hand Side (RHS) and Slack Variables columns [154]. LHS-RHS represents the terms on the left hand side of inequalities of the equation, and on the right hand side of inequalities, respectively. Slack variables are used to transform inequalities to equalities. Each row of ILP's matrix represents one constraint's equation, while each column of ILP's matrix represents one variable of the equation.

6.4.2 Development

In our satellite network allocation problem, our objective function (5.3.7) becomes the objective function of this ILP problem, which is the sum of the cost on each ISL. Cost on each ISL is simply a multiplication of the amount of flows in that ISL and the cost per unit flow on that link. In addition to this, we introduce our privilege parameter into the objective function of this ILP problem, which minimizes the delay time and maximizes the remaining bandwidth capacity. The LP solver in our simulation is based on [155]. In addition to that we use the constraints in (5.3.12, 5.3.18, 5.3.19), as the constraints of this ILP problem.

From table 6.4.1 the constraint equations are given in each row. For example in the first row, the constraint has coefficient 1 for $Path1$, coefficient 1 for flows in $path2$, and coefficient 12 for

Table 6.4.1: LP-matrix sample

LHS								RHS	Slack Variables			
Cons-	Path1	Path2	Link1	Link2	Link3	Link4	Link5		Var1	Var2	Var3	Var4
traints	1	1	12	15	10	0	0	23	1	0	0	0
equa-	0	1	2	0	0	1	2	20	0	1	0	0
tions	0	1	2	2	2	2	2	12	0	0	1	0
	1	0	1	2	2	4	5	13	0	0	0	1

flows in $link2$. The RHS column represents the term on the right hand side of the inequalities. On the first row, RHS is 23 for the first constraint. $Var1$ in Slack variables has a coefficient value of 1 to transform inequalities to equalities in the first constraint.

In combination with Genetic Algorithm from the previous section, the output of the LP-solver constructs a new alternative solution. If this new solution has a better fitness value than the worst solution in the previous population of Genetic Algorithm, it replaces this worst solution.

6.5 Tabu Search

The main purpose of tabu search is to escape from local optima. This type of search is based on procedures designed to cross boundaries of feasibility or local optimality, which can be seen as a barrier to finding the optimal solution. Tabu search methods explore the solution space beyond the local optimality of all feasible solutions by using a sequence of moves. To prevent cycling, some moves are classified as tabu.

Fred Glover gave the first presentation of Tabu Search in the current form in 1986, followed by his report in 1989 [156]. Many computational experiments such as those conducted in [157, 158] have shown that tabu search has now become an established approximation technique, which can compete with other known techniques. In the first paper of Xu, Chiu et al., tabu search was used to optimize link capacities in a dynamic routing telecommunications network. The authors develop and introduce a probabilistic move selection and coordinated solution recovery strategies. While in their second paper, a problem of finding an optimal degree constrained Steiner tree, whose nodes and edges are weighted by costs, is considered. This problem addresses the problem in designing a private line digital data service (DDS) network

over a finite set of customer locations.

6.5.1 Description

In tabu search, the information relates to the local neighborhood. Information related to the exploration process needs to be kept in order to improve efficiency of the exploration process. Besides keeping information about the current value of objective function in the memory, tabu search maintains information about the history of the last visited solution in its memory. Tabu search moves procedure based on short term and long term memory. Short term memory is used to prevent the search from being trapped in a local optimum. Short term memory restricts the possible moves. How long a given restriction operates depends on a parameter called the tabu tenure. Long term memory is used to achieve a diversification effect and encourages the search to explore regions, which less frequently visited.

According to Glover, memory's role in tabu search is restricting the choice of some subset $S(i)$ of solution i by forbidding moves to some neighbor solutions. The structure of the neighborhood of $N(i)$ varies from iteration to iteration as $N(i,k)$ for iteration k. Tabu search iteration is terminated following some stopping conditions. The termination can happen when k is larger than the maximum number of iterations allowed, or the number of iterations since the last solution's improvement is larger than a specified maximum number, or when there is no unexplored neighborhood solution ($N(i,k+1) = \emptyset$). Otherwise, the iteration will be terminated if evidence can be given that an optimal solution has been found [156].

6.5.2 Development

In our research we do not implement the tabu search itself, but instead we use the properties of tabu search in our handover procedure. The first property is the short term memory property of tabu search, which we use in order to expand our solution space and to deal with the complexity of the handover that need to be done due to the satellite movement. The short memory property is implemented by allowing an overcapacity Interplane Satellite Link not to be included in the available links' table for the next two time intervals (see figure 6.5.1). It means that the tabu tenure time is two time intervals. This procedure provides us with a more distributed traffic load, since the incoming traffic is diverted from an over capacity region to a high remaining

bandwidth capacity region. Another benefit is that less number of links will be processed in a one time interval. Therefore, the processing time can be reduced.

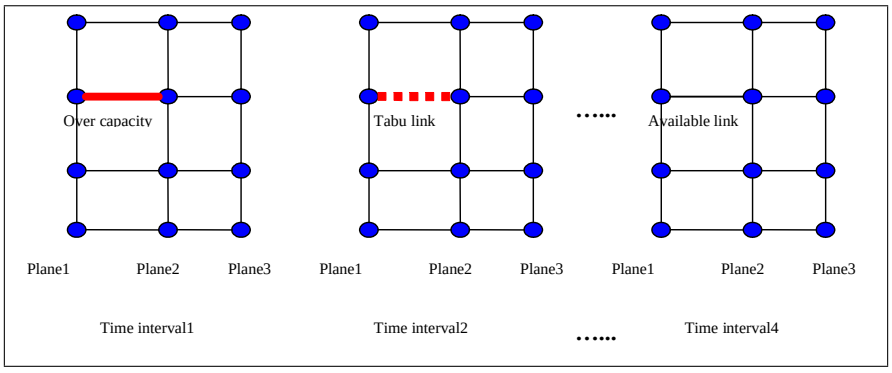

Figure 6.5.1: Short term memory properties of tabu search

6.6 Dijkstra's Shortest Path Algorithm

The shortest path problem is concerned with a problem of finding a path from two given vertices x and y in a weighted graph, with the property that the sum of the weights of all edges is minimized over all such paths. A shortest path algorithm can be used to solve problems including: single-source shortest path problems, single-destination shortest path problems, single-pair shortest path problems, and all-pairs shortest path problems. In single source shortest path problems, we try to find a shortest path from a source s to any destination v in the graph; whereas in single-destination shortest path problems a shortest path to a given destination t from any satellite in graph is solved. A single-pair shortest path will find a shortest path from a source s to destination t, and all-pair shortest path problem will find the shortest paths for every vertices in graph G_r.

Depending on whether weights on the vertices are non-negative or allow a negative value, two different algorithms can be used to solve the shortest path problem. The Dijkstra shortest path algorithm solves the shortest path problem in a non-negative weighted graph, while Bellman-Ford algorithm solves the shortest path problem in either non-negative or negative weighted graph [159].

Shortest path problems occur in many fields. In 1956, Edsger Dijkstra introduced an efficient

algorithm to solve shortest path problems in graphs and demonstrated his algorithm in the ARMAC computer conference. Dijkstra is a Dutch computer scientist and mathematician who has developed many algorithms and his shortest path algorithm is probably the best known [160].

6.6.1 Description

A shortest path problem in a weighted directed graph $G_r = (V, E)$, with weight function $w : E \rightarrow R$ maps these edges to a real valued weight. The weight of $pathP$, which includes vertices: $v_0, v_1, \ldots, v_k$, is the sum of the weights of its edges:

$$w(P) = \sum_{i=1}^{N} w(vi - 1, v_i) \tag{6.6.1}$$

We can define the shortest path weight from source s to destination t by

$$\partial(s,t) = \begin{cases} \min\{w(P) : s \xrightarrow{P} t \quad if \ \ there \ is \ a \ path \ \ from \ \ s \ \ to \ \ t \\ \infty \qquad\qquad\qquad\qquad\qquad\qquad otherwise \end{cases} \tag{6.6.2}$$

Shortest path from satellite s to destination t is then defined as any path P with weight

$$w(P) = \partial(s,t) \tag{6.6.3}$$

Breadth first search can also be interpreted as a shortest path algorithm which works on an unweighted graph. Each edge can be considered as having one unit weight.

6.6.2 Development

A shortest path tree is almost similar to the one from the breadth first search tree, except that it contains the shortest path from the source, which is defined in terms of edge weights instead of number of edges. In our satellite allocation problem, we consider both of these aspects: edge weights and the number of edges, to find the shortest path.

Assume that the satellite constellation can be represented by a weighted graph $G_r : (V, E)$ with weight function $w : E \rightarrow R$; and assume that G_r contains no negative weight cycles

reachable from the source $s \in V$. Then we can define a shortest path tree from s as a directed subgraph $G'_r : (V', E')$ where $V' \subseteq V$ and $E' \subseteq E$, such that:

1. V' is the set of vertices reachable from s in G_r

2. G'_r forms a tree with root s.

For all $v \in V$', we can have a unique path from s to v in G'_r, which is a shortest path from s to v in G_r (the resulted shortest path itself does not need to be unique).

We use (5.3.9) to find the weights (costs) of different paths. Weights of each link depend on two terms: hop lengths (distance between satellites) and residual bandwidth. Consider a problem of finding the shortest path from source s to destination t, with $s, t \in V$. Initially a solution space S contains only source s. On each iteration step, S contains a set of vertices whose final shortest path weights from the source s have already been determined. For all other satellites $i \in V$, we assign a parameter $\pi[i]$ to the predecessor of i. $\pi[i]$ is used to define the adjacency list of an already discovered satellite i. For our graph $G_r = (V, E)$ with source s, we can define the predecessor subgraph of G_r as $G_{r\pi} = (V_\pi, E_\pi)$, where :

$$V_\pi = \{i \in V : \pi[i] \neq NIL\} \cup \{s\} \tag{6.6.4}$$

and

$$E_\pi = \{(\pi[i], i) \in E : i \in V_\pi - \{s\}\} \tag{6.6.5}$$

In addition to the above, we have $d[i]$ as the estimated shortest path from s to i till the current iteration step. $d[i]$ functions as an upper bound on the weight for the shortest path from source s to satellite i. During the execution of this algorithm we repeatedly decrease this upperbound on the actual shortest path weight of each satellite until the upperbound equals to the shortest path weight.

We use the interesting property of shortest path for the handover procedure. Since the subpaths of the shortest paths are shortest paths, given $\mathrm{G}_r = (\mathrm{V}, \mathrm{E})$, assume that there is a demand for a connection between satellite s to destination with a shortest path $p_{s,t} = \{s, i_1, i_2, \ldots, i_k, t\}$ then $p_{i_1,i_k} = \{i_1, i_2, \ldots, i_k\}$ a subpath of p becomes a shortest path from satellite i_1 to satellite i_2. When the satellite topology changes the existing connection is handed over to another

shortest path, which uses the subpath of the previous shortest path and adds a new link to source and destination.

In figure 5.4.5, at time interval 1, a source is serviced by satellite 1 ($sat1$) and the destination is covered by satellite 4 ($sat4$). The connection consists of ISLs: $Sat1Sat2$, $Sat2Sat3$, and $Sat3Sat4$. At time interval 2, the satellite moves in the direction given in the figure 6.6.1. At this time, satellite 5 ($sat5$) covers the source and satellite 6 ($sat6$) covers the destination. Therefore, we handover the connection from $sat1$ to $sat5$, and we add an extra ISL (a new link) between $sat1$ and $sat5$ in this thesis. Moreover, for the destination we insert another additional ISL between $sat4$ and $sat6$. The new path between the source and the destination consists of ($sat5$, $sat1$, $sat2$, $sat3$, $sat4$, *and* $sat6$).

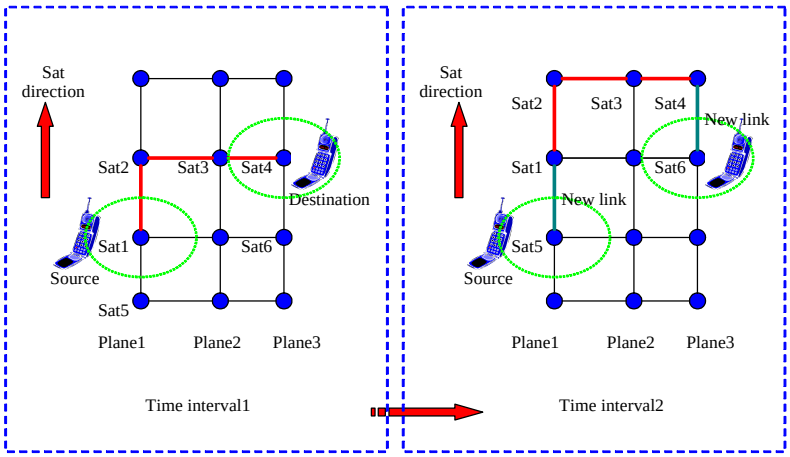

Figure 6.6.1: Handover of the connection from source to destination, by adding additional links from satellite 1 to satellite 5 and from satellite 4 to satellite 6

6.7 Summary

In this chapter we discussed various algorithms which are used in this thesis. Different algorithms perform different tasks and the constructed combination algorithm is beneficial to be implemented in the LEO satellite network environment. GALP performed an optimization of global traffic allocation in the satellite network, which resulted in a more evenly distributed traffic load. In addition to that the privilege parameter provided a method to cope with

multiservice traffic. Furthermore, the Dijkstra shortest path algorithm performed a local optimization of the traffic in the network. The short memory property of tabu search, enhanced the distribution of traffic load over a whole network; and reduced the processing time required to find a solution.

Chapter 7

GALPEDA: GENETIC ALGORITHM LINEAR PROGRAMMING - EXTENDED DIJKSTRA 'SHORTEST PATH' ALGORITHM

7.1 Introduction

In this chapter, a discussion of our combination algorithm Genetic Algorithm Linear Programming and Extended Dijkstra shortest path Algorithm (GALPEDA) is given. GALPEDA is used to solve two problems -the periodical problem and the incremental problem. First, we consider the periodical problem, in which we used GALP. Thereafter, the incremental problem is solved by the EDA part of our GALPEDA algorithm. At the end of this chapter various assumptions, which have been made in order to simulate the use of our GALPEDA in the LEO satellite network, are given. Since we evaluated the performance of GALPEDA to allocate multiservice traffic in the LEO satellite, the parameters that we are evaluating are either from the satellite topology, traffic model, or the GALPEDA itself.

7.2 GALPEDA

The traffic allocation problem in satellite communication is divided into two allocation problems. The user's mobility is negligible relative to the rapid movement of LEO satellites. Therefore, the orbital period of satellites becomes a significant aspect when a connection has to be constructed between the origin and destination. We propose two distinct traffic allocation problems: the periodical problem and the incremental problem. In the periodical problem, we examine the dynamic topology of the LEO satellite network, while in the incremental problem we assume that satellites position remain unchanged. GALPEDA (Genetic Algorithm Linear Programming and Extended Dijkstra shortest path Algorithm) is used to solve these allocation problems.

7.2.1 Periodical Problem

In the periodical problem, traffic demands are constructed from mobile users and dedicated users such as Gateways. The optimal solution is calculated to allocate traffic demand for the whole network. First network topology and network constraints are updated to their initial values at the beginning of each time interval. The positions of the satellites can be predicted in each time interval. We define positions of satellites for a particular time interval following two line elements, which consists of the Keplerian elements. Format of two line elements is given as follow (see Chapter 3):

```
1 NNNNNC NNNNNAAA NNNNN.NNNNNNNN +.NNNNNNNN +NNNNN-N +NNNNN-N N NNNNN
2 NNNNN NNN.NNNN NNN.NNNN NNNNNNN NNN.NNNN NNN.NNNN NN.NNNNNNNNNNNNN
```

As an example for IRIDIUM satellites, we should have for a particular time the following two line elements [120]:

IRIDIUM 8

```
1 24792U 97020A 02034.35939857 -.00000134 00000-0 -54814-4 0 7365
2 24792 86.4010 144.2719 0002455 49.3052 310.8353 14.34215101248708
```

We define the satellite plane position using the second element of Keplerian element, with column number:

03-07, which identifies satellite number

09-16, which gives the inclination angle (between the orbital and equatorial plane

18-25, which gives the RAAN (angle, measured at the center of the earth from the vernal equinox to the ascending node).

Satellite positions are given as a SatPos matrix ($Nx3$), with their latitude and longitude position:

$$\begin{bmatrix} SatelliteNumber & Latitude & Longitude \\ 1 & \theta_1 & \varphi_1 \\ 2 & \theta_2 & \varphi_2 \\ ... & ... & ... \\ N & \theta_N & \varphi_N \end{bmatrix} \tag{7.2.1}$$

We assume that there are M zones of satellite coverage on the earth; we also assume that a connection from MTs will be build inside these zones. Instead of building per MTs connection, we build a connection based on zones. Zones matrix position can be given as a ZonePos $(M \times 3)$ matrix:

$$\begin{bmatrix} ZoneNumber & Latitude & Longitude \\ 1 & \phi_1 & \psi_1 \\ 2 & \phi_2 & \psi_2 \\ ... & ... & ... \\ M & \phi_M & \psi_M \end{bmatrix} \tag{7.2.2}$$

We assume that there are $l_{i,j}$ parallel ISLs from satellite s_i to satellite s_j , with $i, j \in \{1..N\},\ i \neq j$. Total bandwidth of these $l_{i,j}$ parallel ISLs will be $b_{i,j}$ which is the maximum capacity of ISL between two satellites, satellite s_i to satellite s_j. Traffic demand from zone i to zone $j, i, j \in \{1..M\},\ i \neq j$ is characterized by an MxM matrix $D_{zone}(t)$. We denote $d_{zone,i,j}(t)$ as a traffic demand from zone i to zone j.

$$D_{\text{zone}}(t) = \begin{bmatrix} d_{zone,1,1}(t) & d_{zone,1,2}(t) & d_{zone,...,...}(t) & ... & d_{zone,1,M}(t) \\ d_{zone,2,1}(t) & d_{zone,2,2}(t) & d_{zone,...,...}(t) & ... & d_{zone,2,M}(t) \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ d_{zone,M,1}(t) & d_{zone,M,1}(t) & d_{zone,...,...}(t) & d_{zone,...,...}(t) & d_{zone,M,M}(t) \end{bmatrix} \tag{7.2.3}$$

For a system with N satellites in M zones, zone's demand matrix $D_{zone}(t)$ needs to be mapped to $D_{sat}(t)$, satellite's demand matrix. Traffic demand in satellite k is constructed of zones $\{i \ldots j\}$, with

$$i, j \in \{1..M\}, i \neq j; k \in \{1..N\}; \{i, ..., j\} \subseteq SatCoverage_k \tag{7.2.4}$$

SatCoverage is the area inside a satellite footprint on earth. Zone i is in satellite coverage k, if $d_{up/downlink,i,k}$ is

$$d_{up/downlink,i,k} = \left\{ \min_{1 \leq l \leq N} \left\{ d_{up/downlink,i,l} \right\}; \max_{1 \leq l \leq N} \left\{ b_{up/downlink,l} \right\} \geq Demand_{up/downlink} \right\} \tag{7.2.5}$$

Zone i will use satellite k, only if $b_{up/downlink,i,k}$, residual capacity for the up/down link of satellite k for zone i can accommodate the up/down link demand of zone i $Demand_{up/downlink,i,k}$ and only if $d_{up/downlink,i,k}$ is the distance between zone i and satellite k is the shortest distance with $k \in \{1..N\}$, and $i \in \{1..M\}$.

Distance of zone i to satellite k, $d_{up/downlink,i,k}$ is given as (see figure 7.2.1)

$$d_{up/downlink,i,k} = h_{sat}^2 + 4R_{earth}^2 - 2R_{earth}^2 \cos(\theta - \phi) - 2R_{earth}^2 \cos(\varphi - \psi) \tag{7.2.6}$$

(θ,φ) gives the inclination and the RAAN of the satellite respectively; (ϕ,ψ) give the latitude and longitude of the MTs zone respectively. Apply (7.2.6), and with respect to (7.2.5), we can select a satellite, which can accommodate the demand from the corresponding zone. Therefore, we can map zone demand matrix $D_{zone}(t)$ into the satellite demand matrix $D_{sat}(t)$, which is an $N \times N$ matrix

$$D_{\mathrm{Sat}}(t) = \begin{bmatrix} d_{sat,1,1}(t) & d_{sat,1,2}(t) & d_{sat,...,...}(t) & ... & d_{sat,1,N}(t) \\ d_{sat,2,1}(t) & d_{sat,2,2}(t) & d_{sat,...,...}(t) & ... & d_{sat,2,N}(t) \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ d_{sat,N,1}(t) & d_{sat,N,1}(t) & d_{sat,...,...}(t) & d_{sat,...,...}(t) & d_{sat,N,N}(t) \end{bmatrix} \tag{7.2.7}$$

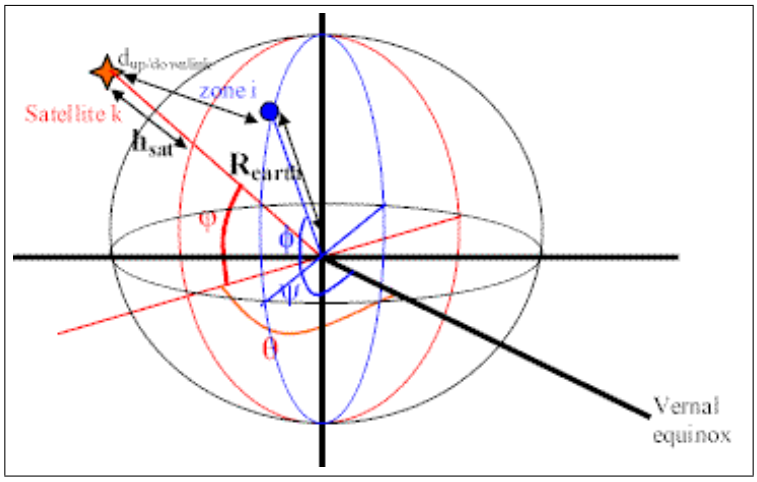

Figure 7.2.1: distance between satellite k and zone i

The demand $d_{sat,k,l}$ from satellite k as source and satellite l as destination is constructed by the demand from all zones $d_{,zone,i,j}$, which have their source in the coverage of satellite k ($satFootprnt_k$), and satellite l covers their destination ($satFootprnt_l$).

$$d_{sat,k,l} = \sum_{i,j \in \{1,M\}; i \in satFootprnt_k; j \in satFootprnt_l} d_{zone,i,j} \tag{7.2.8}$$

Genetic Algorithm and Linear Programming is used to allocate traffic demand into satellite links. The procedure is as follows. At the beginning we generate an initial population, which satisfies the current constraints and demands by using a heuristic algorithm based on the shortest path. The current constraints are given as before, and the following equation is a constraint for the upperbound of traffic using node n, which will include the traffic with destination n, $i_{k,n}$; traffic from n as source, $i_{n,k}$; and traffic, which uses node n as intermediate node, b_{nl}.

$$\frac{1}{2}(\sum_{k=1}^{N}(\nu_{k,n} + \nu_{n,k}) + \sum_{n \neq l} b_{nl}) \leq Q_n \tag{7.2.9}$$

We consider that all traffic demand has a positive value and is below the maximum capacity of the ISLs.

$$0 \leq v_{k,l} \leq Q_{kl} \tag{7.2.10}$$

In our research, we represent chromosomes as a collection of flow paths, with each path carrying a corresponding quantity of traffic from the satellite's source to its destination.

For example, Chromosome:

$FlowpathP_{1,2,3,5}$: 3 $units\ of\ traffic\ flow$

$FlowpathP_{1,2,3}$: 2 $units\ of\ traffic\ flow$

$FlowpathP_{1,2,5}$: 1 $units\ of\ traffic\ flow$

$FlowpathP_{2,3,5}$: 3 $units\ of\ traffic\ flow$

Etc.

In addition, we add information regarding the remaining capacity: the time that a particular link is overcapacity (to consider this link as a tabu link), and the type of traffic that occupies this link.

In our thesis, we use only an empty space at the start time of the simulation. In the following time intervals, we use the previous time interval solution as the first set of solutions in the initial population. The previous solution is updated according to the satellite's new position. The procedure is as follows:

```
Repeat until the problem is solved or no feasible solution can be found
Begin
    Set C as current cheapest cost to infinity;
    For each satellite source s, with s∈{1,N} do
    Begin
        Find chain using Depth First Search to carry demand
            from this satellite, which satisfies the objective
            function with privilege parameters for different traffic types;
        If C is a cheaper path then update C;
        Add to solution;
    End;
    Update remaining ISL and satellite capacities;
End;
```

Then using the concept of Genetic Algorithm, we try to optimize the fitness of this population through the recombination and mutation of the genes. Each chromosome encodes the set

of paths that carry traffic in a particular solution. We start GALP from the initial population, which we achieved from the previous heuristic. The procedure is as follow

```
Repeat n times
Begin
    Randomly choose two parent chromosomes
        from initial population and one extra chromosome
        from the modified solution of the previous time interval;
        Form a union of these two parent chromosomes and one extra chromosome;
        Use LP-solver to allocate traffic and find a new child (solution);
    If the fitness value of the new solution is better than
        the worst chromosome in the current population
        replace this otherwise do nothing;
End;
```

We transformed the union of two parent chromosomes to the input format of LP-solver. Crossover is performed in the LP-solver. We try to optimize the fitness of this population through recombination and mutation of their genes. Two parent chromosomes are chosen from the current population and one extra chromosome from the previous solution is used at each stage. A union is formed from these two parent chromosomes and one extra chromosome. By using LP, we allocate the traffic on this enlarged set of paths. This produces an optimal child. If the new solution is better than the worst solution in the initial population, the new solution replaces the worst solution. If that is not the case then there should be no change in the initial population. At the end of this procedure cycle, the current solution is written in a path directory, which will be used as a starting point to solve the incremental problem. This path directory will be used for the whole period of time interval. At the end of the time interval, after updating the satellite topology (satPos matrix updating) we perform the same procedure to find the best solution for the next time interval.

A different type of crossover procedure has been studied, which results in further complexity. Therefore, we decided to consistently follow the previous procedure. It is useful to discuss these different procedures. Since the number of satellites in the satellite constellation can be huge, to reduce the computational time we consider dividing the problem space z into two sub spaces,

z_1 and z_2. We try to solve the problems in these two sub spaces separately and at the end we combine all of these sub spaces (figure 7.2.2). The demand between these two sub spaces will be introduced by allowing the border satellites on each sub spaces to function as a Gateway between these two sub spaces.The following equation defines the subspace problems

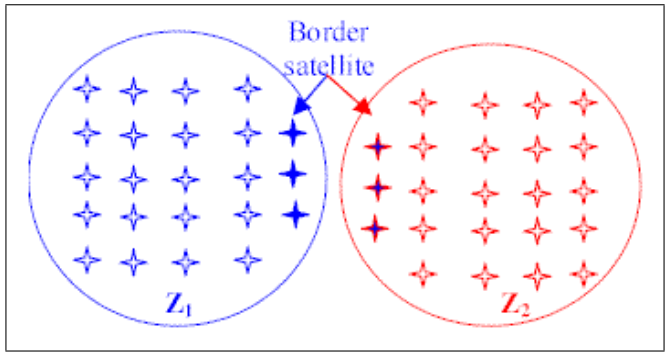

Figure 7.2.2: Alternative solution using subspace

$$\max_{i,j\in\{1,N\}} Z_{i,j} = \max_{i,j\in\{1,\frac{N}{2}\}} Z_{1,i,j} + \max_{i,j\in\{\frac{N}{2}+1,N\}} Z_{2,i,j} + \max_{i\in\{z_1 border\},j\in\{z_2 border\}} Z_{i,j} \qquad (7.2.11)$$

In this subspace problem, we maximized the total objective function of $Z_{i,j}$ (with i and j the satellites in the network of N satellites) by adding the two sub spaces $Z_{1,i,j}$ and $Z_{2,i,j}$. The first two terms in (7.2.11) give the objective function of these two sub spaces. The first one gives the objective function of Subspace Z_1 of satellite 1 to satellite $N/2$, while the second term gives the objective function of Subspace Z_2 of satellite $(N/2+1)$ to satellite N. The last term in this equation is the objective function of the bordering satellites. The processing time to solve this subspace problem is shorter compared to the processing time to solve the whole space problem at once. On the other hand, a more complicated computation is needed when the inter sub spaces connection is required and the border satellites sometimes become overloaded.

Different types of mutation procedures have been studied in our research to diversify the solution space and to escape from local optima. The first mutation procedure is undertaken by adding an extra path, which is not available in the current solution into the input matrix of the LP-solver. We can search an extra path from the previous solution, which is not in the

current solution to reduce searching time. The probability that a mutation occurs is given as the mutations probability. The second mutation procedure is by adding an extra chromosome into the input matrix of the LP-solver. Instead of using two chromosomes three chromosomes are used. The third chromosome originates from the un-repaired solution of the previous time interval.

First, we choose two parent chromosomes from the current population at each stage to perform a crossover. A union is formed from these two parent chromosomes and by using the LP-solver we allocate the traffic on this enlarged set of paths.

The objective function of LP is simply to minimize the total cost in the satellite network, which is given as the sum amount of traffic on each link multiplied by the cost per-unit flow on that link (see(5.3.9)).

The constraints for this LP are demand constraints (the number of connections between satellites, sources to destinations), ISL capacities (maximum traffic which can be accommodated on each link), and satellite capacities (the maximum traffic which can be accommodated on this particular satellite).

Considering these constraints, the LP-solver allocates traffic demand between the satellites and tries to minimize the cost of the whole network. As a result the LP-solver produces an optimal child. If the new solution is better than the worst solution in the initial population, the new solution replaces the worst solution. If that is not the case, then there is no change in the initial population. The current solution is written in a path directory, which is used as a starting point for solving the incremental problem.

7.2.2 Incremental Problem

In this incremental problem, the network topology is assumed to be unchanged. When a subsequent call arrives as an origin destination (OD) pair, then a search for an available path in the path directory is initiated to satisfy this OD pair demand. If there is no available path in the path directory which can satisfy this demand, then Extended Dijkstra shortest path Algorithm is initiated to find a new path. The new path is added into the path directory. The current solution from the path directory is inserted into the new initial population for the next periodical problem, after upgrading the paths in the directory by predicting the new positions

of the satellites.

Since we are dealing with the demand of various traffic classes, then we introduce a mechanism in which more privilege is given for higher priority traffic. In order to do this we introduce a parameter λ^c into our problem formulation. λ^c is used as given in (5.4.15). Since we assume that the solution paths for a subsequent request will be processed only at the source, we route the requests using source routing. Each satellite maintains a path directory database that contains a description of the satellite network state as known to that satellite. Information about the residual bandwidth will be stored in an $N \times N$ residual Bandwidth matrix B_{res}. The ISL's state database on each satellite will be updated by distributing satellite network state information using the flooding method. We only update the whole ISL's state database when the change of the remaining capacity in ISL is below a certain delta threshold value $\Delta b_{threshold}$ or when it reaches the end of the time interval. Otherwise, we update only information about the changing of ISL's state of one particular satellite. An event initiates a new link state update when the available bandwidth in a link changes significantly. We assume that b_{last} is the residual capacity at the last update, and $b_{current}$ is the current residual capacity then an update notification is flooded only when

$$\frac{|b_{current} - b_{last}|}{b_{last}} > \Delta b_{threshold} \tag{7.2.12}$$

In the second term in (5.4.15) we introduce a notional cost based on the unused capacity of the transmission link. As available bandwidth of a transmission link is limited, it is important to share traffic over the whole network. When all links have a spare capacity, then the ability to handle additional demand is enhanced. Hence, when adding a new demand to an existing pattern of traffic, we assign a higher cost if the traffic is allocated to a link with little remaining capacity. This tends to divert new traffic to paths with spare capacity. The two terms in objective function may sometimes conflict, since the minimum hop route will not necessarily minimize the cost due to the bandwidth available. By introducing λ^c in (5.4.15) we give a preference to a high priority class of traffic and provide a mechanism to give preference to lightly-loaded links and allow us a better control of traffic distribution. We use a large value of λ^c for a low priority class of traffic, and a small value of λ^c for a high priority class of traffic.

By assigning a large value of λ^c to low priority class of traffic, this class of traffic is diverted to lightly-loaded ISLs. Hence, we can reserve some amount of bandwidth for high priority traffic. The choice of the value λ^c depends on how Big the reserve space should be given for the high priority class and also, on the maximum hops allowed to re-route the low priority traffic. In our research, we consider 2 classes of traffic, namely the high priority traffic and the low priority traffic. Therefore, we have two values of λ^c namely λ^{high} and λ^{low} for high priority and low priority traffic respectively. λ^c values depend on the maximum acceptable hop length diversion between the classes, and the threshold capacity on all of the links $b_{threshold}$. The minimum value of $\lambda^{low}/\lambda^{high}$should be chosen as:

$$(\frac{\lambda^{low}}{\lambda^{high}})_{\min} = \frac{(\triangle\ n(P_{kl}^{high}, P_{kl}^{low}))}{\sum\limits_{(i,j)\in P_{kl}^{high}} \frac{1}{b_{ij}^{threshold}} - \sum\limits_{(i,j)\in P_{kl}^{low}} \frac{1}{b_{ij}^{threshold}}} \tag{7.2.13}$$

in which:

$\Delta n(P_{kl}^{high}, P_{kl}^{low})$ is the max acceptable total hop difference in path P_{kl}^{high} for high priority and P_{kl}^{low} for low priority traffic.

$b_{ij}^{threshold}$ is the threshold capacity between link i to j.

P_{kl}^{high} is $((k, i_1), (i_1, i_2), \ldots, (i_{l-1}, l))$ a shorter path from k to l, which is a sequence of links for high priority traffic

P_{kl}^{low} is $((k, j_1), (j_1, j_2), \ldots, (j_{l-1}, l))$ a longer path from k to l, which is a sequence of links for low priority traffic

Consider satellite constellation in figure 7.2.3, suppose that there is a demand request for a connection from *satellite1* to *satellite*3, with two available paths. The first path is a shorter *pathI* ($P_{1,3}$) and the second one is a longer *pathII* ($P_{1,2,4,3}$).

We use equations (5.3.3) and (5.3.4) to determine the path length for the first path and the second path. For the sake of simplicity, we assume that the distances between satellites are one unit length. When we use *pathI* using (5.3.16) and consider that the time needed to flow for

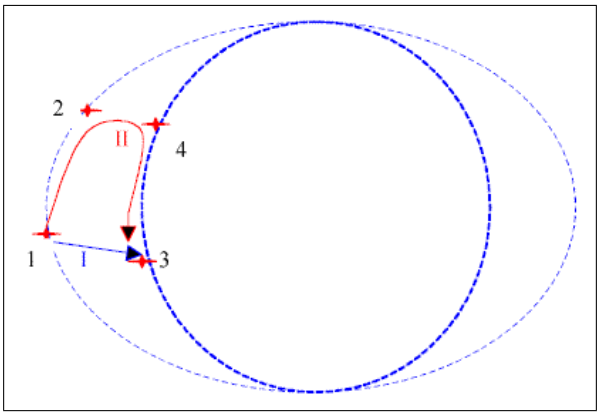

Figure 7.2.3: Two available alternative paths

one unit length of ISL is one, then the objective function from (5.4.15) becomes:

$$\min(\sum_{k=1}^{N}\sum_{l=1}^{N}(\sum_{i=1}^{l}(((T_{up/down}+t(P_{kl}))+\frac{\Gamma}{b_{k,l}})v_{i,kl}))) \tag{7.2.14}$$

With assumption that the up and downlink time is a constant $T_{up/down}$ and the upper term of residual penalty cost in (5.3.16) as a constant Γ. Then the cost of using *path I* ($z_{total(I)}$) with one unit demand becomes:

$$z_{total(I)} = (T_{up/down}+1)+\lambda_I^c\frac{\Gamma}{b_{1,3}} \tag{7.2.15}$$

while the cost for *pathII* ($z_{total(II)}$) becomes:

$$z_{total(II)} = (T_{up/down}+3)+\lambda_{II}^c(\frac{\Gamma}{b_{1,2}}+\frac{\Gamma}{b_{2,4}}+\frac{\Gamma}{b_{4,3}}) \tag{7.2.16}$$

The incoming demand with source *satellite1* and destination *satellite3* uses *pathII*, when:

$$\frac{\lambda_{II}^c}{\lambda_I^c} = \frac{(T_{up/down}+3)-(T_{up/down}+1)}{\frac{\Gamma}{b_{1,3}}-\left(\frac{\Gamma}{b_{1,2}}+\frac{\Gamma}{b_{2,4}}+\frac{\Gamma}{b_{4,3}}\right)} \tag{7.2.17}$$

Therefore, we can divert the low priority traffic to a more lightly-loaded link, even though the path could be longer, when we choose $\lambda^{high}$and λ^{low} properly, according to (7.2.13). This

approach allows us to distribute the traffic more evenly over the whole network. By choosing a high value of λ^{low} and low value of λ^{high}, bandwidth reservation can be achieved for high priority traffic.

In the following example we show a simple case of a satellite network with 9 satellites in 3 planes (see figure 7.2.4). We use EDA to allocate two classes of traffic, which will differ a low priority class of traffic to a lightly-loaded traffic. We assume that at time $t = T_0$ the satellite

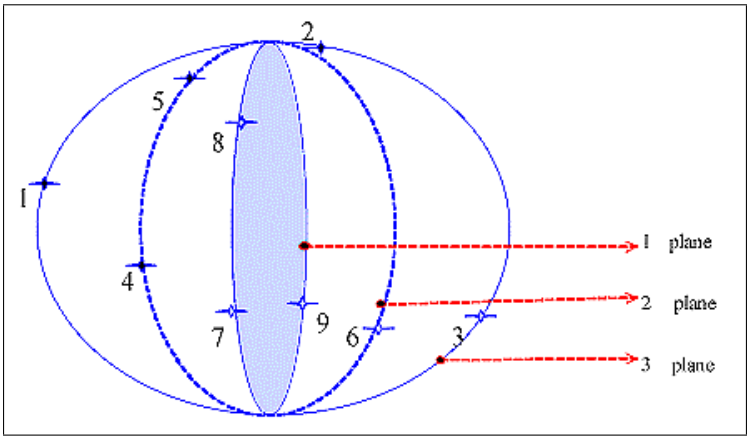

Figure 7.2.4: Satellite constellation with 9 satellites in 3 planes

demand D_{Sat} and the available bandwidth b_{Sat} matrix are

$$D_{Sat}(T_0) = \begin{bmatrix} 0 & 8 & 1 & 8 & 7 & 6 & 5 & 7 & 7 \\ 8 & 1 & 7 & 6 & 5 & 6 & 5 & 7 & 6 \\ 1 & 7 & 2 & 4 & 3 & 0 & 4 & 4 & 5 \\ 8 & 6 & 4 & 2 & 4 & 0 & 5 & 6 & 7 \\ 7 & 5 & 3 & 4 & 2 & 6 & 6 & 6 & 7 \\ 6 & 6 & 0 & 0 & 6 & 2 & 5 & 6 & 6 \\ 5 & 5 & 4 & 5 & 6 & 5 & 2 & 7 & 6 \\ 7 & 7 & 4 & 6 & 6 & 6 & 7 & 2 & 7 \\ 7 & 6 & 5 & 7 & 7 & 6 & 6 & 7 & 2 \end{bmatrix}, b_{Sat}(T_0) = \begin{bmatrix} 10 & 2 & 9 & 2 & 3 & 4 & 5 & 3 & 3 \\ 2 & 9 & 3 & 4 & 5 & 4 & 5 & 3 & 4 \\ 9 & 3 & 8 & 6 & 7 & 10 & 6 & 6 & 5 \\ 2 & 4 & 6 & 8 & 6 & 10 & 5 & 4 & 3 \\ 3 & 5 & 7 & 6 & 8 & 4 & 4 & 4 & 3 \\ 4 & 4 & 10 & 10 & 4 & 8 & 5 & 4 & 4 \\ 5 & 5 & 6 & 5 & 4 & 5 & 8 & 3 & 4 \\ 3 & 3 & 6 & 4 & 4 & 4 & 3 & 8 & 3 \\ 3 & 4 & 5 & 3 & 3 & 4 & 4 & 3 & 8 \end{bmatrix} \quad (7.2.18)$$

We choose for this example $\lambda^{low} = 15$ for low priority traffic and $\lambda^{high} = 1$ for high priority

traffic. At time $t = T_1$, a low priority call arrives with demand 1 from source satellite 1 to destination satellite 4. Then, at time $t = T_2$ a high priority call arrives with demand 2 from the same source-destination pair satellite. The Dijkstra shortest path algorithm attains the shortest path for the incoming demand request using the objective function ($min\ z_d$), while EDA assigns a different value of λ_c for a different priority class and uses the objective function in (5.4.15).

We consider firstly the shortest path algorithm case. At time $T_{0.}$, the available bandwidth $b_{Sat}(T_0)$ between satellite 1 and 4 is equal to 2 ($b_{1,4}(T_0)$=2). At time $t = T_1$ when a low priority request arrives with demand 1 the loaded traffic is allocated in path $(1, 4)$. The available bandwidth in $b_{1,4}(T_1)$ reduces to 1. At time $t = T_2$, if another high priority request arrives with demand 2 and maximum hop length 1 between satellite 1 and 4, there is no available bandwidth between $(1, 4)$ which can satisfy this high priority demand, since the shortest path $b_{1,4}(T_2)$ has only 1 available bandwidth remaining. Thus, the incoming high priority request is blocked.

We consider now the case with the EDA algorithm. At time $t = T_1$ we have two choices of path from the source satellite 1 to the destination satellite 4. The first path is $p_{1,4}$. This has the shortest time delay and an alternative path is $p_{1,3,6,4}$. When a low priority traffic request arrives at time $t = T_{0.}$ with demand 1 the cost of using path $p_{1,4}$ is:

$$z_{p1,4} = (T_{up/down} + 1) + \lambda^{low}(\frac{\Gamma}{b_{1,4}}) \tag{7.2.19}$$

if we use path $p_{1,3,6,4}$. total cost is as follows:

$$z_{p1,3,6,4} = (T_{up/down} + 3) + \lambda^{low}(\frac{\Gamma}{b_{1,3}} + \frac{\Gamma}{b_{3,6}} + \frac{\Gamma}{b_{6,4}}) \tag{7.2.20}$$

We suppose that the up and downlink propagation time $T_{up/down}$ is similar for both cases since they use the same source-destination satellite and we use λ^{low}= 15, and Γ=1, then cost of using $path_{1,4}$ is 8.5 ($c_{1,4} = 1 + 15 \times \frac{1}{2}$) and cost of using path $p_{1,3,6,4}$ is 7.67 ($c_{1,3,6,4} = 1+15\times\frac{1}{9}+1+15\times\frac{1}{10}+1+15\times\frac{1}{10}$). Since at time $t = T_0$ we have $b_{1,4}(T_1) = 2$, $b_{1,3}(T_1)$ =9, $b_{3,6}(T_1)$ =10, $b_{6,4}(T_1)$ =10. Because the cost for the second path is lower than the first path, the second path is chosen for the incoming low priority request. If a high priority request

demand arrives at $t = T_2$ with 2 unit demand and maximum hop length of 1 from satellite 1 to destination satellite 4, we use λ^{high}=*1*. The cost of the first alternative solution, *path* $p_{1,4}$ will be 1.5 ($c_{1,4} = 1 + 1 \times 1/2$) and for the second alternative *path* $p_{1,3,6,4}$ will be $4\frac{25}{72}$ ($c_{1,3,6,4} = 1 + 1 \times \frac{1}{8} + 1 + 1 \times \frac{1}{9} + 1 + 1 \times \frac{1}{9}$). Therefore, high priority traffic is allocated into *path* $p_{1,4}$.

This multi service approach can reserve some bandwidth for the possible incoming high priority traffic, whereas in the previous shortest path algorithm the high priority request is blocked.

In case of very sparse traffic in a satellite constellation, the shortest path algorithm allocates the incoming request into the best available path disregarding the priority of the incoming request. There should be no reservation for the high priority traffic, which means no waste of bandwidth, due to the bandwidth reservation for high priority requests. In EDA some amount of bandwidth is reserved for the high priority traffic. In case the amount of high priority traffic is low, some amount of bandwidth remains unused.

In case of very dense traffic in a satellite constellation, both algorithms perform comparably. We investigate the case of that the available bandwidth matrix at $t = T_0$ is as follows:

$$b_{Sat}(T_0) = \begin{bmatrix} 10 & 2 & 1 & 2 & 1 & 1 & 1 & 1 & 1 \\ 2 & 9 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 8 & 1 & 2 & 1 & 1 & 1 & 1 \\ 2 & 1 & 1 & 8 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 8 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 1 & 8 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 5 & 8 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 8 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 8 \end{bmatrix} \tag{7.2.21}$$

When we use EDA for an incoming request of low priority class, we obtain for the first path: 8.5 ($c_{1,4} = 1 + 15 \times \frac{1}{2}$), and for the second alternative path $p_{1,3,6,4}$ is 33 ($c_{1,3,6,4} = 1 + 15 \times \frac{1}{2} + 1 + 15 \times \frac{1}{2} + 1 + 15 \times \frac{1}{2}$).

In the EDA we use a priority queue Q, which contains all satellites in a satellite network

G, keyed by their current shortest distance values $d[i]$ (table 7.2.1). We represent the satellite network G as adjacency list. The procedure of the EDA is as follows:

Initialize single satellite source (satellite s); {(G, s)}

Empty the solution space S; {S:= $\varnothing$}

fill the priority queue according their d[i, λ] values; {Q := V(G)}

If class is low then λ:=λ^{low} else λ:=λ^{high};

While Q is not empty do {Q$\neq \varnothing$}

Begin

Extract the nearest satellite u from Q; {U:= Extract min Q }

Include satellite u into the solution space; { S := S$\cup${u}}

For each satellite i $\in$ Adj[u] do

Begin

If d[i, λ]> d[u] +weight[u,i, λ] then

π[i, λ] $\longleftarrow$ u; {Include u into π[i, λ]}

End;

End;

Table 7.2.1: Layout of Dijkstra's shortest path algorithm table

steps	d[sat$_{\mathbf{1}}$]	d[sat..]	d[sat$_{\mathbf{1}}$]	π[sat..]	π[sat..]	π[sat$_{\mathbf{N}}$]	S	V	Q
1	∞	∞	∞	NIL	NIL	NIL	$\varnothing$	Sat$_1$,sat... ,sat$_N$	Sat$_1$,sat... ,sat$_N$
2									
k	δ[sat$_1$]	δ[sat$_1$]	δ[sat$_1$]	π[sat$_1$]	π[sat$_1$]	π[sat$_1$]	Sat$_1$,sat... ,sat$_N$	Sat$_1$,sat... ,sat$_N$	$\varnothing$

The procedure starts by initializing for all satellites an upperbound $d[i]$ which has an infinite value, since no calculation has been done for the shortest path from source s to any satellites in the network. We assign *Nil* for all satellites predecessor $\pi[i]$, which means that no satellite

has an adjacency satellite within a finite distance. At the beginning of the iteration solution space S is an empty space, while the priority Queue Q consists of all the satellites.

7.3 Assumptions and Parameters in The Simulation of GALPEDA

In order to investigate the performance of our combination algorithm, we performed an empirical study using a simulation model. This model carries out several experiments with various satellite network parameters and system parameters. Moreover, the performance of our proposed algorithm to cope with different types of traffic model is examined. Due to the complexity of the problems consideration, some assumptions have been made to reduce the degree of complexity.

7.3.1 Assumptions

In our simulation model, a satellite network is modeled as in Ballard notation with m, the angular distance between adjacent planes.

$$m = \frac{180^0}{P} \tag{7.3.1}$$

with P is number of planes. We do not introduce any spacing offsets in planes of our satellite network. The space segment of our satellite network model consists of cross connect satellites $XCSat$, which have on-board switching capabilities. A constructed connection from the satellite source s to the destination t has only one uplink connection from the source and one downlink connection to destination t. We consider no Gateway connection in between. Instead, we use only ISLs to construct a connection from source to destination. $Total_{up/downlink}$ delay will have a maximum of $(2 \times h_{Sat}/velocity\ of\ light)$ for any connection.

In this thesis, we consider only satellite constellations with circular orbits, which have an inclination of 90^0. In polar regions, we assume that ISLs still have the same degree of intersatellite connection as in equatorial regions, to reduce the complexity of the handover in this region. Furthermore, we do not introduce any seam region into our model for the same reason. Satellites perform as a loss station, with a finite number of ISLs and no waiting room ($M/M/S(0)$ queue model).

The satellite's coverage model is an earth fixed cell model. We do not consider handover

caused by beam handover. We only consider an inter-satellite handover, which includes intra-plane and inter-plane handovers.

We investigate only transmission loss in relation to the blocking of a signal, if there is no available connection to satisfy the incoming demand. We do not take into account the attenuation loss, and Doppler Effect. In addition, we are not considering the mobility of the MT on earth, since the MT speed is relatively much slower than the speed of satellites, as explained in Chapter 4. We can assume that a mobile user on earth is a static object, and the satellite is a moving object. Therefore, the mobility of satellites becomes the main issue in the mobility management of our traffic allocation problem.

Another assumption is due to the traffic type, which is explained in more detail in the following paragraph.

7.3.2 Traffic Models

According to the ITU Recommendation H.261 [85] the video stream is composed by a sequence of frames. MPEG2 codes video and audio according to this standard. Each frame consists of a number of packets, with their responding QoS requirements. Since we consider QoS routing with differentiated services architecture, we classified the incoming traffic into two types of traffic, CBR and VBR/NRT as given in Chapter 2. The two types of traffic, CBR and VBR/NRT are called: *priority 1* for the non-delay sensitive traffic, and *priority 6* for delay sensitive traffic in the traffic class field of IPv6 as given in Chapter 2 [105].

Non-delay sensitive traffic (class of low priority) consists of traffic with VBR/NRT, and is constructed of data traffic and any video playback or video mail messages. This is considered as low priority traffic. This traffic represents asynchronous traffic and has a negative exponential distribution packet length.

Delay sensitive traffic (class of high priority traffic) consists of traffic with CBR and is constructed of voice traffic (with an on and off process) and video traffic (streaming and real time). This class of traffic has a fixed length of packet size. Voice traffic generates a fixed packet size in its talking state, whereas the video traffic generates a fixed packet size in its active state.

The inter-arrival time of these packets is modeled as either a negative exponential distribution (Poisson process) or as a Markov Modulated Poisson Process (MMPP) [12].

The first model is the Poisson traffic model is one variant of renewal traffic models. In such models, interarrival time A_n is IID (Independent, Identically Distributed) and their distribution can be general. Poisson traffic model is a renewal traffic model whose interarrival times are exponentially distributed with rate parameter α, which is

$$prob\{A_n \leq t\} = 1 - \exp(-\alpha t) \tag{7.3.2}$$

Poisson processes have some useful properties for the implementation of our satellite constellation model. Superposition of independent Poisson processes results in a new Poisson process with the rate being the sum of the component rates. This means that we can accumulate the arrival rate in each satellite to get the total arrival rate for the whole satellite network. The independent increment property gives memory-less properties of Poisson processes. Then it is allowable to define a time dependent Poisson process by letting rate parameter α depend on time, known as the Switched Poisson Process (SPP). This will be the next step in future research. Since we would like to vary α with the time of the day, e.g. α will be higher in the day time than at night time. Renewal processes are relatively simple because they are independent of each other, thus the correlation between the arrivals becomes zero. This correlation carries information about the 'burstiness' of the traffic. Auto-correlated traffic models are essential for predicting the performance of emerging broadband networks. Since Poisson carries no information about 'burstiness' of traffic, we also model the interarrival time according to the Markov model.

The second model is a Markov-Modulated Poisson Process (MMPP). Since renewal traffic models have their drawbacks (they can not be used to transfer information about the 'burstiness' of traffic), we use a Markov traffic model. This model introduces a dependency random sequence of inter-arrival times $\{A_n\}$, which can potentially capture 'burstiness' information. Let M be a Markov process with a discrete state space

$$M = \{M(t)\}_{t=0}^{\infty} \tag{7.3.3}$$

The inter-arrival times $\{A_n\}$ depends on the state from where the jump occurs, as it depends on transition matrix $X = [x_{ij}]$ We use one variant of the basic Markov model, Markov Modulated

Poisson Process (MMPP). This model is almost similar to SPP. The difference is that in SPP transitions the rate is in addition dependent of t (subscript state) transitions, in which it combines the simplicity of modulating the Poisson process into the Markov Process. In MMPP model, for each state k of M, arrivals occur according to a Poisson process at rate α_k, and when the state changes, so does the rate. In general we can define a matrix T as a transition matrix of the modulating Markov chain and Λ as a matrix whose diagonal elements contain the arrival intensities that corresponds to the different states of the chain, as given in [162]:

$$T = \begin{bmatrix} -\omega_{11} & \omega_{12} & ... & \omega_{1m} \\ \omega_{21} & -\omega_{22} & ... & \omega_{2m} \\ ... & ... & ... & ... \\ \omega_{m1} & -\omega_{m2} & ... & -\omega_{mm} \end{bmatrix} \tag{7.3.4}$$

$$\Lambda = diag(\alpha_1, \alpha_2, ..., \alpha_m) \tag{7.3.5}$$

In our simulation model we consider that the modulation Markov chain contains only two states, with ϖ_1, ϖ_2 and α_1, α_2.

7.3.3 Parameters

The first performance parameter that we investigate is the performance of our proposed algorithms in various satellite constellation parameters. In our simulation analysis, we investigate performance of our algorithm in a various number of satellites, a various number of planes, and a various latitude of satellites orbits.

We investigated the performance of algorithms in various arrival rates and interarrival times between incoming packets.

The duration of voice traffic when it is at on state and video traffic when it is in active state is investigated as well. The longer the average duration of this traffic, the more handover processing needs to be done.

Moreover, the performance of algorithms with two different types of traffic model is investigated. The first traffic model - the Poisson based traffic model will not represent the 'burstiness' of the current traffic, while the Markov based traffic model can represent the 'burstiness'.

7.3.4 Simulation Model

We model our simulation using event scheduler controls. This event scheduler controls the simulation's clock; scheduled activities are ordered chronologically by the scheduled time of their occurrence. The simulation clock is updated to the time of the next event. In this approach, we use the critical event approach. A State changes when an event occurs. Three events cause these states' transitions, which are:

- Arrival of a new request: A new request is generated with an exponential distribution with rate λ and mean value μ (in traditional traffic model). After arrival of a new request the state will change.

- Departure of a request: Duration of a request with CBR (voice in on state and video in active state) is uniformly distributed. When a request is ended then it is removed from the channel.

- Updating period: Satellite topology is updated at the beginning time intervals. In which existing traffic load is reallocated following the new topology.

In our simulation, traffic is modeled in packet level, in which each request is represented in a small packet. Voice traffic has one unit length of packet in its on-state (similar to video traffic in its active state). Since routing is considered in the network layer, a packet header is used as given in figure 7.3.1:We consider two traffic classes with source and destination addresses that represent their satellites number. We use the same hop limit for all requests.

At the beginning of simulation time, we start with an initial zone demand matrix for both types of traffic. Location of source and destination are generated randomly with $\phi \in \{0^0, 180^0\}$ and $\psi \in \{0^0, 360^0\}$, while the positions of satellites are generated randomly with $\theta \in \{0^0, 180^0\}$ and $\varphi \in \{0^0, 360^0\}$.

For CBR traffic types we generate one unit of traffic with its duration of request. During this time, we generate one unit of traffic per unit time. While, for the VBR traffic we generate an exponentially distributed traffic load, which will have values between one to ten units of traffic per unit time.

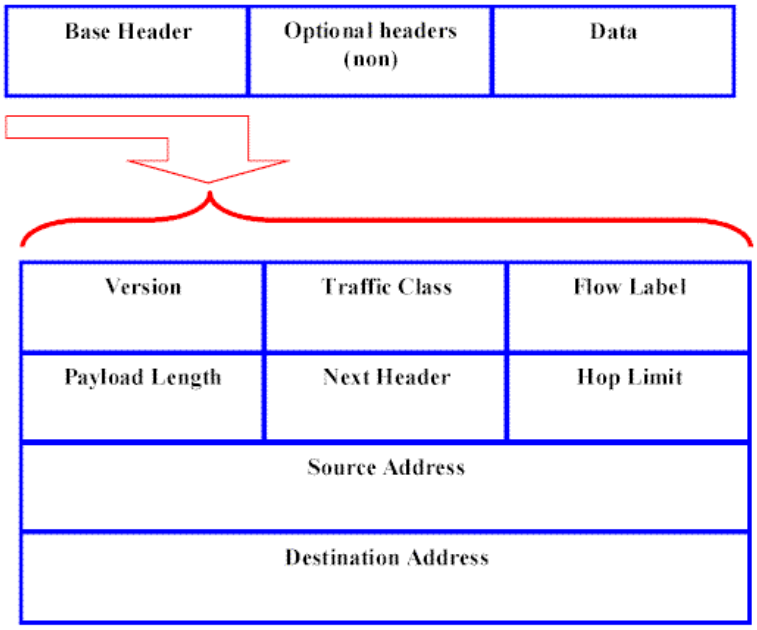

Figure 7.3.1: Packet header

Interarrival times between packets are modeled according to two different traffic models: Poisson and Markov based traffic models. In the first experiment we model interarrival times following the Poisson traffic model as given in (7.3.2) while in the second experiment time interarrival times are modeled between arrivals as MMPP.

Remaining capacities on ISLs in the satellite network construct $N \times N$ available bandwidth matrix $b_{available}(t)$.

$$b_{available}(t) = \begin{bmatrix} b_{1,1}(t) & b_{1,2}(t) & b_{.,.,.}(t) & ... & b_{1,N}(t) \\ b_{2,1}(t) & b_{2,2}(t) & b_{.,.,.}(t) & ... & b_{2,N}(t) \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ b_{N,1}(t) & b_{N,2}(t) & b_{.,.,.}(t) & .... & b_{N,N}(t) \end{bmatrix} \tag{7.3.6}$$

with the corresponding $N \times N$ cost matrix $C(t)$ due to delay is:

$$C(t) = \begin{bmatrix} c_{1,1}(t) & c_{1,2}(t) & c_{..,..}(t) & ... & c_{1,N}(t) \\ c_{2,1}(t) & c_{2,2}(t) & c_{..,..}(t) & ... & c_{2,N}(t) \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ c_{N,1}(t) & c_{N,2}(t) & c_{..,..}(t) & ... & c_{N,N}(t) \end{bmatrix} \tag{7.3.7}$$

At the beginning of simulation time, we generate an initial demand for both traffic types. This traffic demand is allocated into the ISLs and stores the solutions into a path directory, which has the format as (table 7.3.1):

Table 7.3.1: Path directory format

Congestion bit	Origin satellite	Destination satellite	Min capacity	ISL with min capacity	Path Length	Congested time	Path
0	Sat $_1$	Sat_5	8 unit	Sat $_2$, Sat_3	3	1	Sat_1, Sat_2, Sat_3, Sat_5
. .	. .	. .	. .	. .	. .	. .	. .
1	Sat $_5$	Sat $_7$	1 unit	Sat $_5$, Sat $_7$	1	2	Sat $_5$, Sat_7

The path directory stores the current solutions, which can be used for the next request when there is an available capacity. The congestion bit will notify whether the corresponding path is almost saturated (0 is not congested). The origin and destination satellite represents the source and destination pair of satellites, which can use this path. Information about the residual capacity for the corresponding path is given in the fourth field. Both the amount of residual capacity and ISL which have this bottleneck are given. Information about the path length is also given, followed by the satellite's path itself.

Since our simulation model is based on a critical event approach, a stack of critical event times is used (figure 7.3.2). This stack consists of the times that a particular event needs to be processed. There are three types of critical events, arrival of a new request (*event*#0), departure of a request (*event*#1) and a periodic updating (*event*#2). When a request arrives (*event*#0),

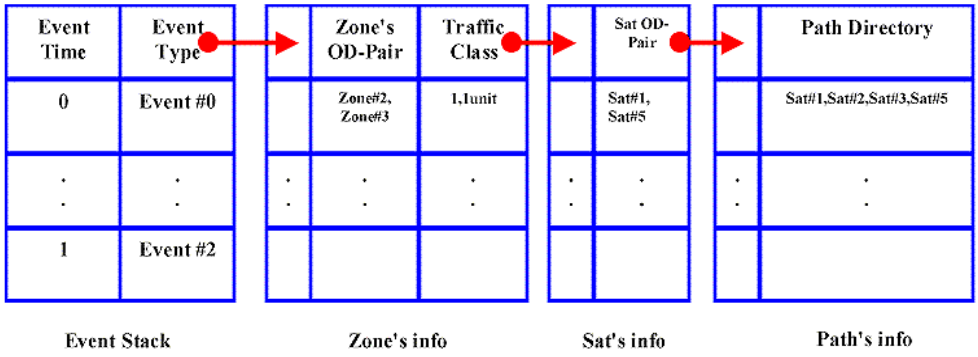

Figure 7.3.2: Events Stack with links to information of zone's/satellite's source destination pair, and their paths

then it generates corresponding subsequent events. Depending on the duration of the request (e.g. for CBR-traffic class), it schedules the departure event of this request (*event*#1). At the same time, it generates the arrival time of the next request depending on the traffic model used, either with SPP or MMPP interarrival times. Periodic updating event (*event*#2) is generated by the simulator.In both events, arrival and departure, a link list with information over the OD pair of zones/satellites is provided. This information consists of a class of traffic, amount of request demand, and the allocated path.

Based on this simulation environment, the performance of our proposed algorithms is investigated. The simulation model performs using Delphi 3 software in a Pentium III 550 MHz with 128Mb of RAM.

7.4 Summary

In this chapter we discussed our proposed new combination algorithm GALPEDA which is used to solve a traffic allocation problem in a LEO satellite network. Each algorithm performed a different task. In this thesis GALP is used to solve a periodical problem and EDA is used to solve the incremental problem. We introduced a privilege parameter, which is used to give a preference to high priority traffic and to distribute traffic more evenly over the whole network. There is a lower bound of the value of this privilege parameter to be able to divert the low priority traffic into a longer path, in case of limited remaining capacity. While a high priority

traffic is given a reserve capacity in a low remaining capacity link. The amount of this reserve capacity depends on the value of our privilege parameter. We introduced an adaptive reserve capacity which can be located for high priority traffic, and which depends on the traffic load on a particular time. We used a revised chromosome from the previous time interval's solution to perform the crossover, in order to decrease the processing time. Instead of starting with an empty space of initial population, we used this revised solution as the first chromosome in the initial population. A mutation effect is added into the GALP by using a non-revised previous time interval's solution to perform the crossover, in order to escape from the local optima. Some assumptions have been made in order to analyse the performance of our proposed routing algorithm when solving a routing allocation problem in LEO satellite network. Assumptions in regard to satellite topology, the mobility of mobile user on earth, and the traffic model have been made.

Chapter 8

SIMULATION OF TRAFFIC ALLOCATION IN LEO SATELLITE USING GALPEDA

8.1 Introduction

In this chapter, we investigate the capability of our GALPEDA to allocate traffic according to its type, to distribute the traffic load more evenly and to determine whether high priority traffic is given more privileges than low priority traffic. Various system parameters of LEO satellites are provided to investigate their effect on the performance of our GALPEDA [150], including the number of satellites and the number of planes in LEO satellite constellation. The Poisson and MMPP traffic models are used in this simulation; and the effect of the various arrival rates of these traffic models is investigated. We compare the performance of our GALPEDA with the previous Genetic Algorithm and Linear Programming of Berry, Murtagh et al. and Montgomery [152,163], which we notates as Genetic Algorithm-Linear Programming 1 (GALP1) in this thesis.

8.2 Simulation Model

The difference between GALP1 and our GALPEDA is given in the following figures.

Shown below is the flow chart of GALP1 (see figure 8.2.1). We modified this flow chart

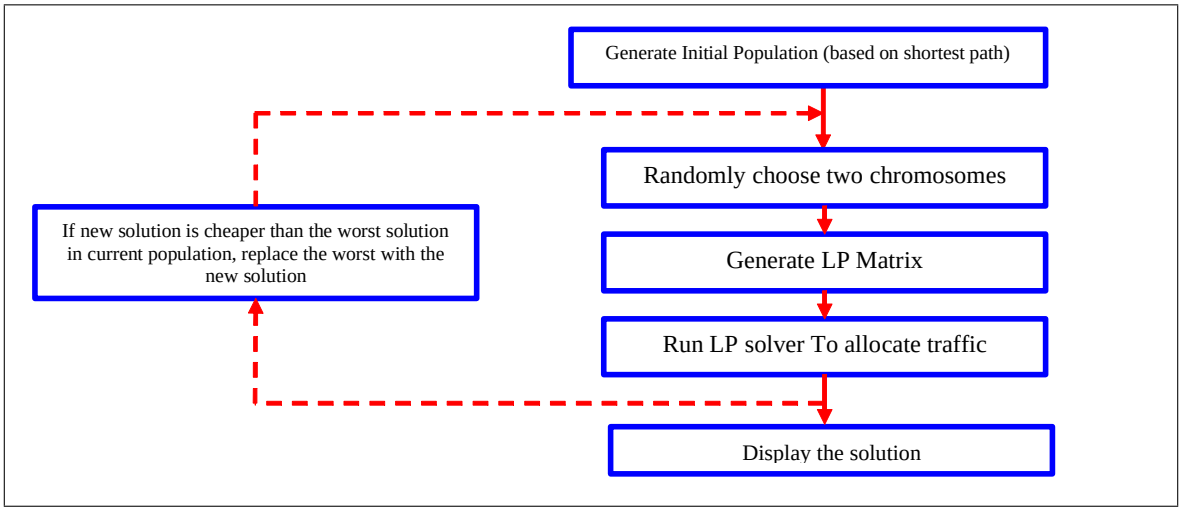

Figure 8.2.1: Flow-chart of GALP1

to cope with satellite constellation parameters as shown below (figure 8.2.2) In GALP1 (figure 8.2.1), the initial population is generated by using a heuristic algorithm to find the minimum hop between the OD-pair. In figure 8.2.2, we use EDA with a privilege parameter to construct the initial population. The initial population in our GALPEDA consists of the collection of flowpaths instead of a collection of paths. In addition to that, we divide the dynamic traffic allocation problem into two separate problems: the periodical problem and the incremental problem. The first occurs when the event is a periodic event, and the second occurs when the event is an arrival event. The first problem is solved with GALP and the second problem is solved with EDA. The handover procedure is added in the periodical problem to update the previous time interval solution. Handover occurs only in the periodical problem, since we make an assumption that the satellite topology changes only at the beginning of the periodical problem. The procedures of our simulation model of GALPEDA, which is based on figure 8.2.2 , are given below:

Begin

Generate satellite positions

Generate initial user's positions and their demand

Map initial user position demand to initial satellite demand

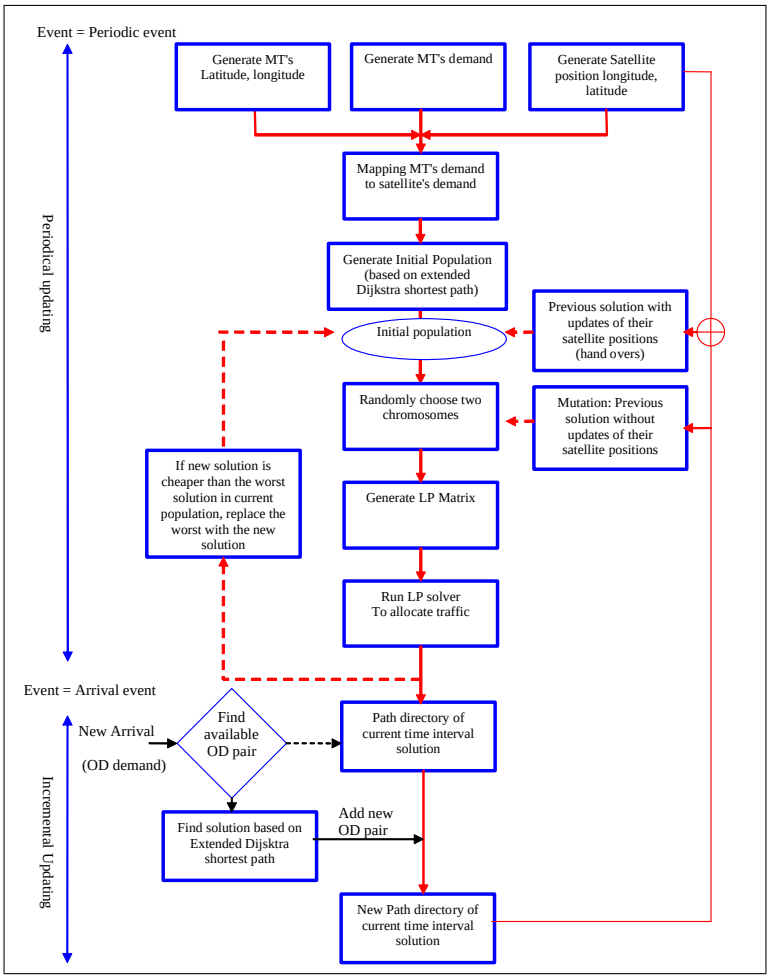

Figure 8.2.2: Flow-chart of GALPEDA

While not (Orbital period of LEO satellite network) do

Begin

If (Event=periodic Event) then

Begin

Upgrade topology;

Upgrade the current solution from path directory;

Generate initial population;

Insert solution from path directory into the initial population;

Repeat 40 times

```
            Begin
                Perform GALP;
             End;
            Save current solution into path directory;
        End
        Else
        Begin
            If (Event=Arrival Event) then
            begin
                Generate a new demand;
                Find OD pair for the new demand;
                If there is an available OD pair then use this
                or else perform EDA and insert new OD pair in path directory;
            end
            or else {Event=Departure Event} remove the ending call;
        End;
End; End;
```

8.3 Simulation Results

Using our model we studied the performance of our GALPEDA under various parameters of GALPEDA, satellite constellation parameters, as well as two traffic models: the Poisson and the MMPP. The duration of simulation was one orbital period of approximately 100 minutes. The simulation was repeated for each case

8.3.1 Performance of GALPEDA with Various Parameters of GALPEDA

Population size

Several simulations were conducted. The first simulation examined the influence of variation in the population size of Genetic Algorithm to the progress speed. A satellite network with 20 satellites and hop-limit of 5 ISLs is considered for this first test. The progress speed performance

is measured by observing the relative bias value. This is a normalized cost difference between the current solution in progress and the final solution, compared with the cost difference between the initial solution and the final solution:

$$relativebias = \frac{C_{progress} - C_{final}}{C_{init} - C_{final}} \tag{8.3.1}$$

figure 8.3.1 shows that GALPEDA progresses rapidly to the final solution by using a small

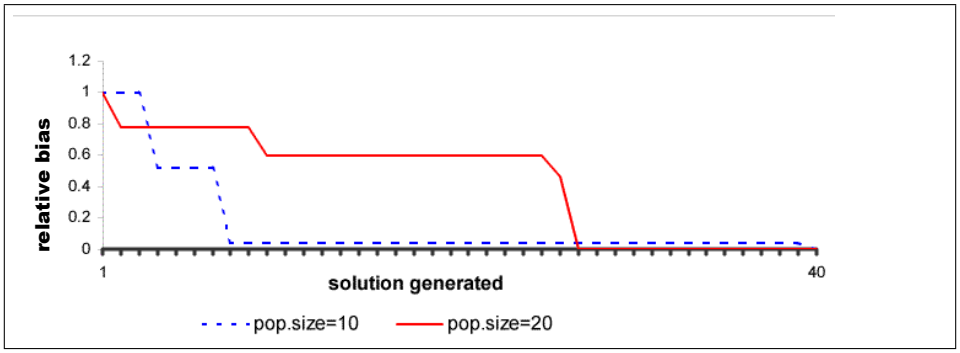

Figure 8.3.1: Relative bias value with different size of population

population size of 10 solutions. However, a small error remains till almost the end of the process. With a bigger population size of 20 solutions, the final solution can be reached after 25 steps. Since rapid progress to a near optimal is significant to guarantee a prompt solution when the topology is updated, a smaller population size is more favorable. In the case of a better accuracy is needed then a bigger population size is necessary.

Hop limit

The second simulation investigates the effect of a hop-limit in the progress speed performance of GALPEDA. In this test, we consider a satellite network with 50 satellites and a population size of 20. figure 8.3.2 illustrates the progress of GALPEDA to reach the final solution with various values for the hop-limit. A similar normalized cost, as given in (8.3.1), is used for the value of relative bias. If a hop-limit of 10 is used, the solution approaches the final value more slowly than for a hop-limit of 20 because of the difficulty of finding a shorter solution for some requests. If a higher hop limit is given, more alternative paths can be given.

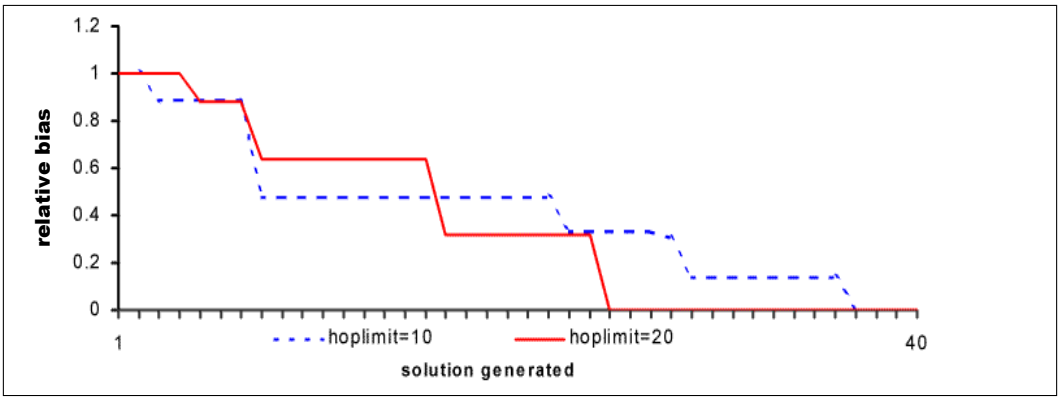

Figure 8.3.2: Relative bias value with two different values of hop-limit

Node degree

In the next simulation, our model does not have a node degree limit and the population sizes 10, 15, and 20 are considered. In this case, there is no limit to the number of ISLs which can be connected to each satellite. We investigate the node degree frequency distribution (the number of satellites which have the same node degree or the same number of ISLs) in different sizes of population.

Figure 8.3.3 shows that when the population size is 10, more satellites have either 4 or 10 connected ISLs to themselves. When the population size is 20, GALPEDA tends to allocate more distributed ISL connections. Each satellite has ISL connections between 5 to 11 ISLs with average node degree frequency of 2. By increasing the population size, GALPEDA can have a more distributed node degree allocation than using a smaller population size, since more alternative paths will be available.

Relative improvement

The fourth test examined the relative improvement of GALPEDA with an increased number of satellites. We use the same initial population from a heuristic algorithm as the starting point for each process. We start with 3 satellites and increase to 30 satellites, with population size

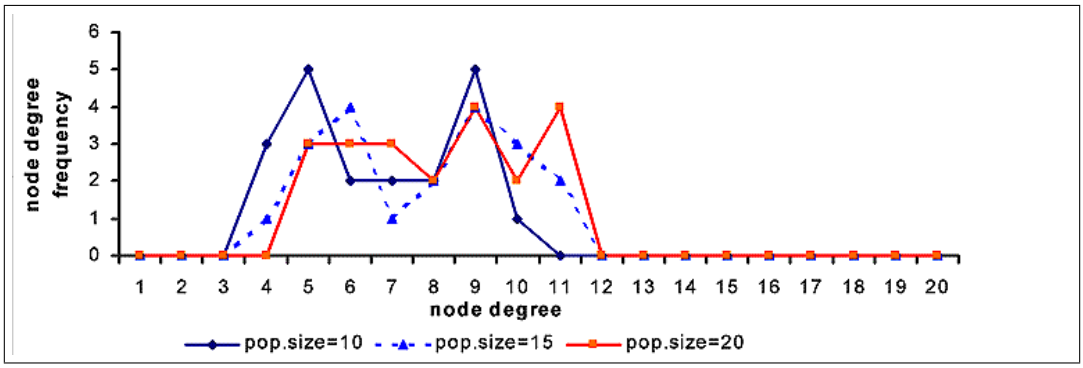

Figure 8.3.3: Node degree frequency distribution with various size of population

of 20 and a hop-limit is 5. The relative improvement for this case is defined as:

$$relativeimprovement = \frac{C_{init} - C_{final}}{C_{init}} \tag{8.3.2}$$

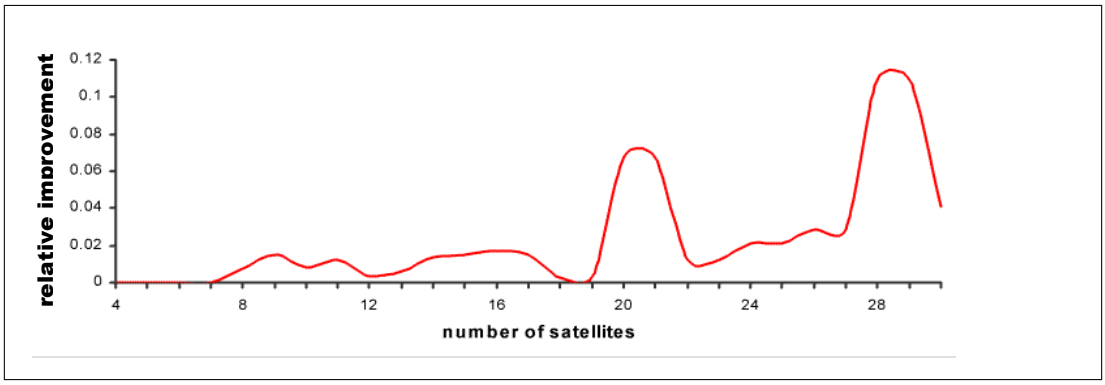

Figure 8.3.4: Relative improvement as the number of satellite is increased

A population size of 20 and a hop-limit of 5 are considered for this test. Figure 8.3.4 illustrates the relative cost value improvement due to the GALPEDA algorithm, by increasing the number of satellites. When the number of satellites is smaller than 8, GALP does not make any significant improvement to the initial cost. Which means that heuristic algorithm produces a near optimal cost value for a small number of satellites. In this case, our GALPEDA cannot

locate a superior solution than this initial solution. In contrast, however with a larger number of satellites, GALPEDA enhances the best cost value at the end of the simulation period. In this figure, we observe that in some cases the improvement is remarkable. A good improvement can be achieved when the number of satellites is 21 (with 3 planes and 7 satellites on each plane) and 28 (with 4 planes and 7 satellites on each plane), since more satellites in one plane provide a more stable connection. This is due to a permanent intraplane ISL connection and the assumptions made for the handover procedure. In the case that a connection should be handovered, the connection will be given to the next satellite in the same plane.

8.3.2 Performance of GALPEDA with Various Parameters of a Satellite Constellation

We investigate the performance of our GALPEDA algorithm starting with various numbers of satellites. Simulation proceeds for a period of one orbital period of LEO satellites (which is assumed to be 100 minutes real time) as follows.

Various numbers of satellites

We commence with a fixed pattern of existing requests, which partly loads the network and use an initial traffic demand. GALP is used to allocate the current traffic on the network. We generate additional traffic requests, which have a Poisson model of interarrival times. These additional requests have an arrival rate of 20 (3 arrivals per second), with the duration of each CBR packet, exactly 10 seconds. Each request is assigned a priority, depending on whether it is CBR or NRT/VBR traffic; a source and a destination. These values all come from uniform probability distributions. The EDA is used to allocate each additional request. Values $\lambda^{low} = 15$ and $\lambda^{high} = 1$ are used for requests of low and high priority respectively. The length of each allocated path is recorded.At the end of each period, the traffic load on each of the ISLs is recorded in four ranges: 25% loaded, 25% to 50% loaded, 50% to 75% loaded, 75% to 100% loaded. Link loading is averaged at the end of simulation. Simulation runs were designed to test effectiveness of the parameters in diverting low priority traffic on to lightly-loaded links, with various numbers of satellites. We consider satellite constellations with 10 satellites to 50 satellites, which are positioned in their 5 orbital planes. In table 8.3.1, the minimum, average

and maximum values of length of path for low and high priority traffic are given. These values as given in this table show that by increasing the number of satellites, the class of high priority traffic has been given a privilege over the class of low priority traffic. The average path length of low priority traffic with 10 satellites is 2.9, while the high priority traffic has an average of 2.8. A satellite constellation with 50 satellites has the average path length of 3.8 and 2.7 for low priority and high priority traffic class, respectively.

Table 8.3.1: Path lengths of low and high priority traffic as the number of satellite is increased

Number of satellites	Number of planes	Low Priority traffic			High Priority traffic		
		min	avg	max	min	avg	max
10	5	2	2.9	5	2	2.8	4
15	5	2	2.9	5	2	2.8	4
20	5	2	2.9	4	2	2.8	4
25	5	2	2.9	5	2	2.8	4
30	5	2	2.9	5	2	2.8	4
35	5	2	3.6	4	2	2.9	4
40	5	2	3.8	6	2	2.7	4
45	5	2	3.8	6	2	2.7	4
50	5	2	3.8	6	2	2.7	4

Furthermore, we analyze traffic distribution in the same satellite environment, as given in figure 8.3.5. This figure shows how the loaded traffic is more evenly distributed in the satellite constellation with the larger number of satellites, while number of ISLs which carry a traffic load of more than 75 % decrease. This is due to the parameter λ, which tries to divert traffic with low priority to lightly-loaded links.

Various numbers of planes

Similar observations have been done for various numbers of planes in the satellite constellation. We consider a satellite network with 24 satellites in 3, 4, 6, and 8 planes. At the beginning, the average length of paths for different priorities of traffic is considered. Table 8.3.2 shows that the number of planes does not have an effect on average length of paths. The reason for this is

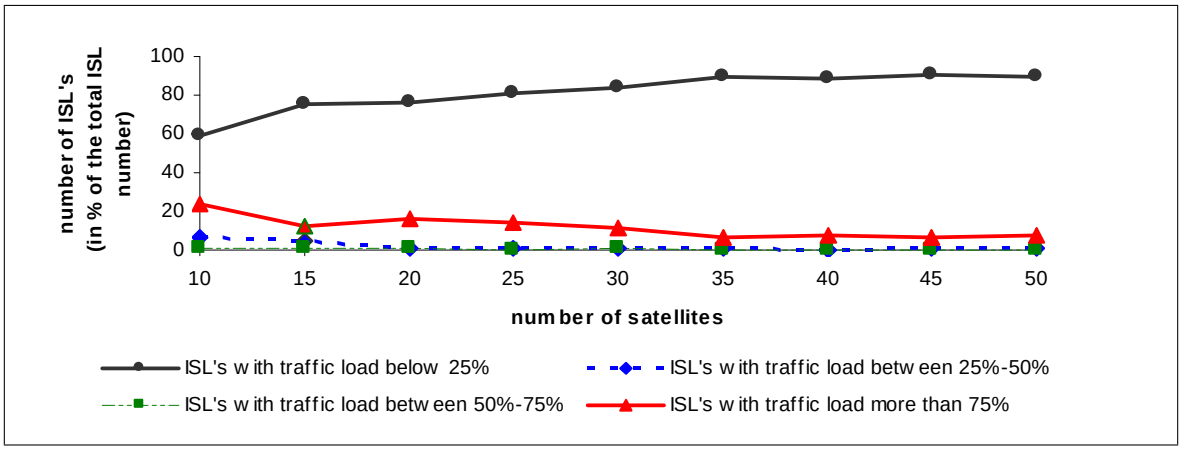

Figure 8.3.5: Traffic load distribution as the number of satellite is increased.

that we consider the same cost value for an ISL between two satellites in the same plane, and between two satellites in neighboring planes.

Number of satellites	Number of planes	Low Priority traffic			High Priority traffic		
		min	avg	max	min	avg	max
24	3	2	2.9	5	2	2.8	4
24	4	2	2.9	5	2	2.8	4
24	6	2	2.9	4	2	2.8	4
24	8	2	2.9	5	2	2.8	4

Table 8.3.2: Path lengths of low and high priority traffic by increase number of planes

However, an investigation of traffic distribution shows that traffic is more evenly distributed when the number of planes is increased. This phenomenon is shown in the following figure 8.3.6. This is a result of our assumption that handover of a connection will only consider neighboring satellites in the same plane. By increasing the number of planes, a connection can have more flexibility in assigning the new connection.

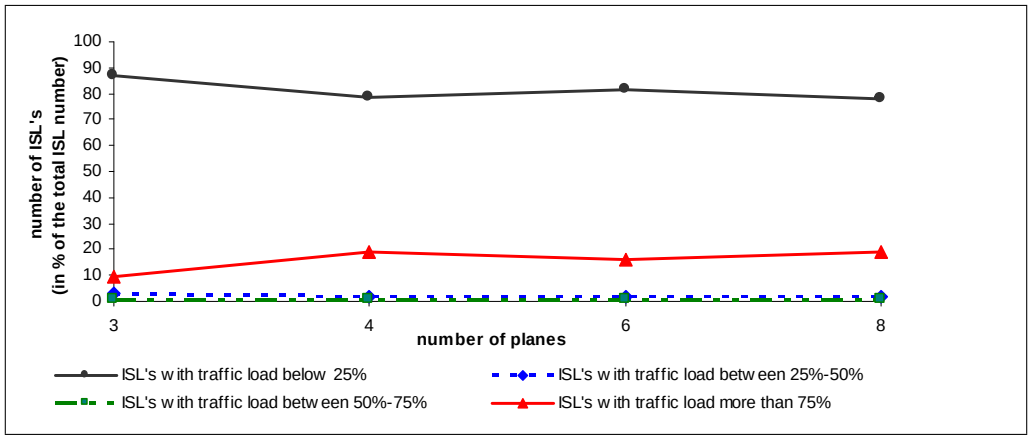

Figure 8.3.6: Traffic load distribution by increase number of planes.

8.3.3 Performance of GALPEDA with Various Arrival Rates

In the next experiment, we consider a simulation of a satellite constellation with only 16 satellites in 4 planes. This is smaller than a real system, which would require at least 48 satellites for global coverage. However, it is adequate to test our model and algorithms.

In our model we assumed that the class of low priority consists of delay insensitive traffic. This traffic represents asynchronous traffic and has a negative exponential distribution packet length with a mean value of 3Kbit. In contrast, the class of high priority traffic consists of delay sensitive traffic. This traffic is constructed of voice traffic (with an on and off process) and video traffic (streaming and real time). This class of traffic has a fixed length of packet size. We assume that the fixed length of this packet size is 3 Kbit. Voice traffic generates a fixed packet size in its talking state, whereas the video traffic generates a fixed packet size in its active state.

Firstly, we investigate the performance of the GALPEDA algorithm by varying the arrival rate of incoming traffic through 20, 40, 60, 80 and 100. λ^{low} is chosen as 15.

Figure 8.3.7 shows that the algorithm performs with more discrimination when the arrival rate is higher. If the arrival rate is 20, the difference between the average path lengths for high and low priority traffic is smaller than when the arrival rate is 100. When more requests arrive, the algorithm allocates the low priority traffic into more lightly-loaded links, to reserve some amount of bandwidth for high priority traffic.

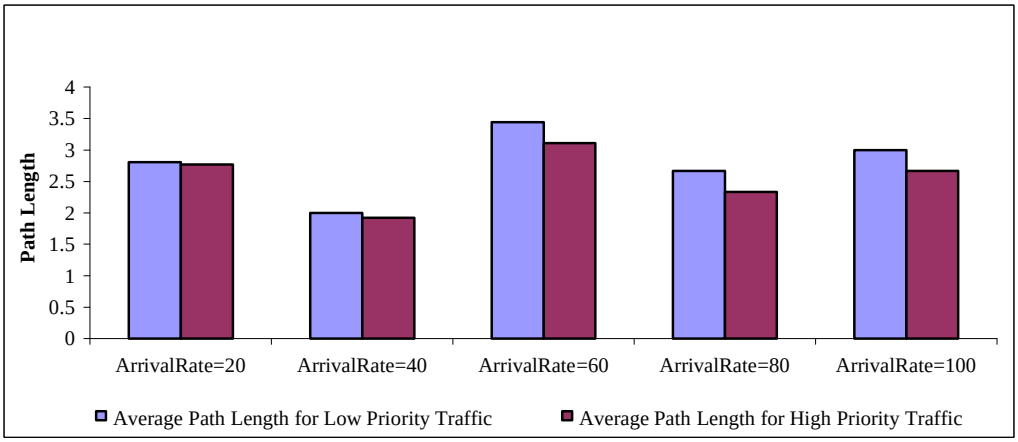

Figure 8.3.7: Average Path length with various Arrival rate

Moreover, we observed call blocking probability of both low and high priority traffic with the increased number of call arrivals (call arrival rate).

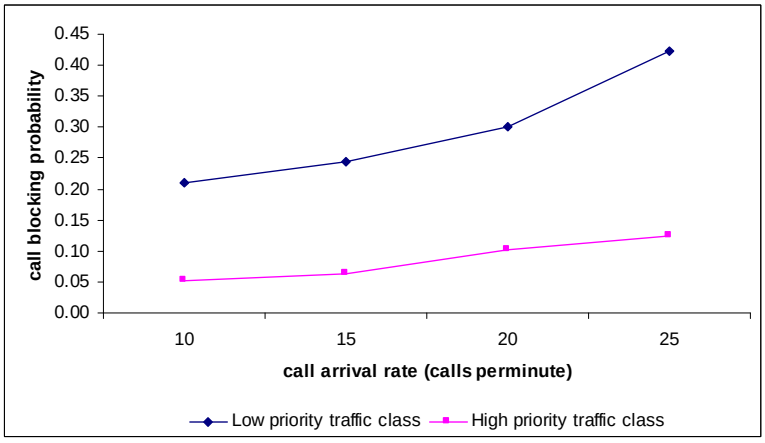

Figure 8.3.8: Call blocking probability of low and high priority traffic class, with a various call arrival rate and the traffic model is a Poisson traffic model

In this experiment, we consider an incoming call into a satellite network, which cannot find an available path between its OD-pair, to be a blocked request or a blocked call. Figure 8.3.8 shows the call blocking probability of the Poisson traffic model, when the arrival rate increases from 10 calls to 25 calls perminute. The call blocking probability of high priority traffic class is

approximately 15% lower than the call blocking probability of low priority traffic. This is due to the privilege that we have been given for high priority traffic class.

Figure 8.3.9 shows the call blocking probability of the MMPP traffic model, when the arrival rate increases from 10 calls/minute to 25 calls/minute. In this case, call blocking probability of high priority traffic class is approximately 17% lower than the call blocking probability of low priority traffic. In case of 'bursty' traffic as modeled in MMPP traffic model, GALPEDA can cope better than in the case of Poisson model traffic, which is due to variable equal length periodic interval. The equal length interval adapts to the change of the incoming traffic.

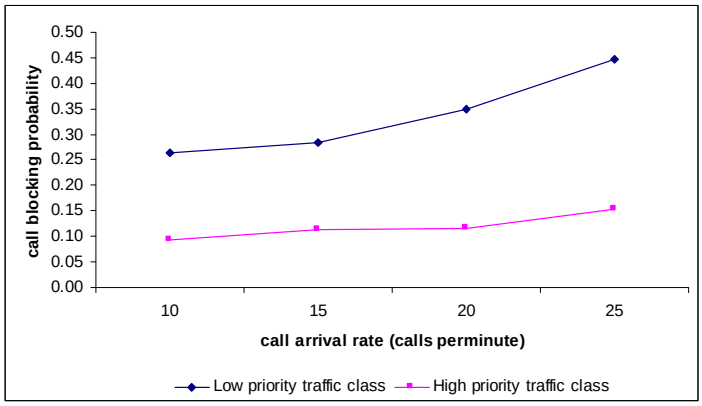

Figure 8.3.9: Call blocking probability of all traffic class in Poisson and MMPP traffic model

If the call arrival rate increases, the difference between call blocking probability of high and low priority traffic class becomes bigger in both traffic models. This is because more ISLs are reserved for high priority traffic class. Figure 8.3.10 shows that if the call arrival rate increases, the performance of our GALPEDA algorithm for both traffic models becomes more similar. This is because of the adaptivity of the length interval. If the network becomes heavy loaded, the periodical updating performs more frequent. Then the traffic load is distributed globally, which results in more evenly distributed traffic load and less call blocking will occur.

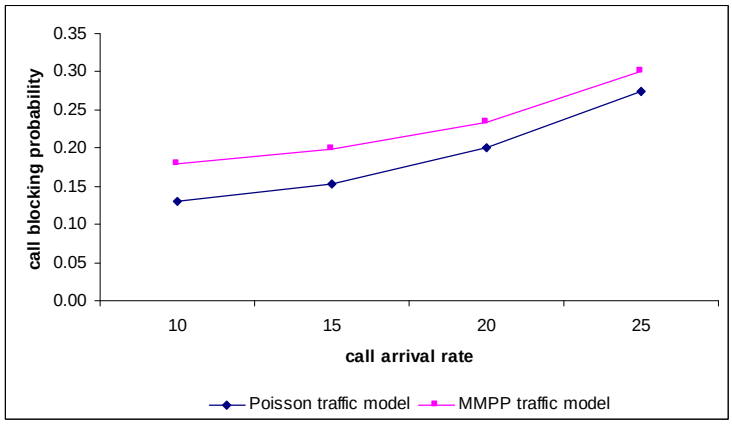

Figure 8.3.10: Call blocking probability of all traffic class in Poisson and MMPP traffic model

8.3.4 Performance of GALPEDA with Two Types of Traffic Model: Poisson and MMPP

Two distinct simulations were conducted, with Poisson and MMPP arrival distributions. In Poisson traffic model, the interarrival times are exponentially distributed with an arrival rate parameter α. In MMPP traffic model, the arrival rate for each state k of M occur according to a Poisson process at rate α_k, and when the state changes, so does the rate. Since we only consider two states in our simulation model, the arrival rate of MMPP at these two states are defined by giving the values of α_1 and α_2, as 0.5 and 1.5 respectively.

As before, the same simulation steps were undertaken. First GALP was used to solve initial demand requests at the beginning of the time intervals. When the subsequent request arrives inside the time intervals, with its arrival rate determined by the traffic model, EDA was used to allocate this subsequent request. Values λ^{low} and λ^{high} were used for requests of low and high priority, respectively. Keeping the simulation value of λ^{high} constant ($\lambda^{high} = 1$), we tested the values $\lambda^{low} = 1$, 5, 10, 15, 20. Simulation was conducted to investigate the effectiveness of the parameter λ^{low} in diverting low priority traffic to lightly-loaded links. First, we consider the Poisson model followed by the MMPP model.

Table 8.3.3 shows that the distribution of the link load varies with λ^{low} when the model is a Poisson model.Table 8.3.4 shows the results when the traffic model is MMPP.

Table 8.3.3: Traffic load distribution for Poisson traffic model

Link Loading	λ^{low} = 1			λ^{low} = 5			λ^{low} = 10			λ^{low} =15			λ^{low} =20		
	Number of Links (%)			Number of links (%)			Number of Links (%)			Number of Links (%)			Number of Link (%)		
	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max
Below 25%	87.6	94.1	100	90.1	94.1	100	88.9	95.2	100	88.9	94.9	100	95.1	96.7	100
25%-50%	0.0	5.2	11.1	0.00	4.8	9.9	0.0	4.5	11.1	0.0	4.9	11.1	0.0	3.2	100
50%-75%	0.0	0.1	2.5	0.00	0.9	5	0.0	0.1	4.9	0.0	0.1	1.2	0.0	0.1	100
75%-100%	0.0	0.3	2.5	0.00	0.2	2.5	0.0	0.3	2.5	0.0	0.1	1.2	0.0	0.1	1.2

Table 8.3.4: Traffic load distribution for MMPP traffic model

Link Loading	λ^{low} = 1			λ^{low} = 5			λ^{low} = 10			λ^{low} =15			λ^{low} =20		
	Number of Links (%)			Number of links (%)			Number of Links (%)			Number of Links (%)			Number of Link (%)		
	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max
Below 25%	85.2	93.6	100	90.1	96.2	100	88.9	94.1	97.0	85.2	94.9	95.2	92.6	97.5	100
25%-50%	0.0	5.5	12.4	0.0	3.4	9.9	0.0	5.5	8.6	0.0	4.4	14.8	0.0	2.1	100
50%-75%	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	2.5	0.0	0.3	2.5	0.0	0.3	100
75%-100%	0.0	0.5	2.5	0.0	0.4	2.5	0.0	0.2	2.5	0.0	0.1	1.2	0.0	0.1	2.5

Both of these tables show that by increasing the value of λ^{low} to 20 we can distribute the traffic load more evenly across the whole network. A higher value of λ^{low} diverts low priority traffic to a longer path, which will decrease the performance of the algorithm.

Furthermore, a significant output of the simulation is path length information. We expect by increasing the value of λ^{low}, that high priority traffic will obtain a shorter path than low priority traffic.It is clearly shown in figure 8.3.11 for the Poisson model and figure 8.3.12 for the MMPP model that by increasing λ^{low}, average path length of high priority traffic decreases whereas the average path length of low priority traffic increases. Since a shorter path for high priority traffic can be provided, this class of traffic will have a lower transmission delay compared to

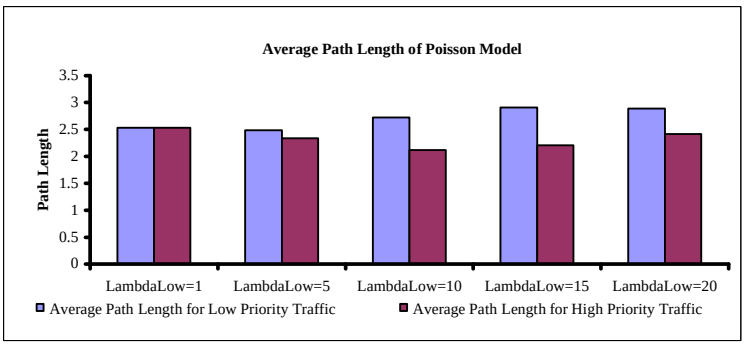

Figure 8.3.11: Average path length of Poisson traffic model

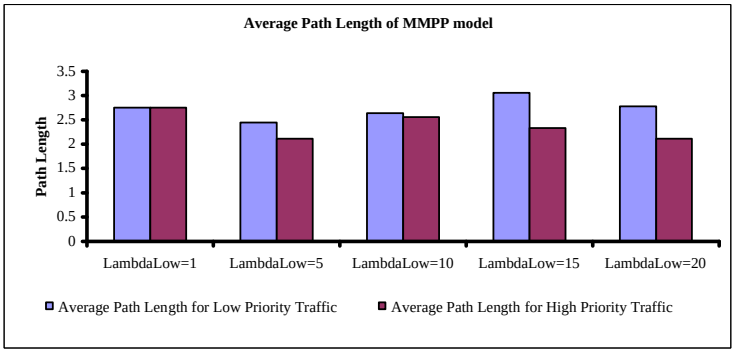

Figure 8.3.12: Average Path Length of MMPP traffic model

the class of low priority traffic. Less number of satellites is necessary to accomplish this shorter path. This reduces the processing time needed to complete the path. This lower transmission delay and less processing time deliver a better QoS for the class of high priority traffic.

8.3.5 Performance of GALPEDA in Average Processing Time

Figure 8.3.13 shows the processing time of our algorithm with various numbers of satellites. We used a 550 MHZ Pentium III machine to obtain this result. The average processing time of GALPEDA increases when the number of satellites is more than 30 satellites. This is the processing time at the beginning of a time interval when an updating of the satellite topology and a traffic allocation for the global satellite constellation occur.

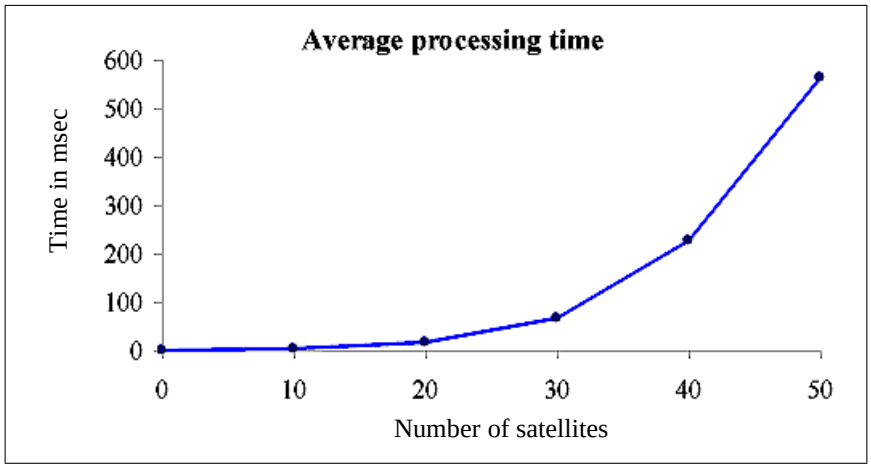

Figure 8.3.13: Average processing time of GALPEDA as the number of satellites is increased

8.3.6 Comparison of GALPEDA with GALP1

In the next case, we compared the performance of GALPEDA with GALP1. We considered in this case 16 satellites in 4 planes, with a hop limit of 10 and the arrival rate of 20 packets per second. The simulation was conducted several times with this system specification, and then the average of traffic load in every ISLs was taken. We classified these ISLs into 4 groups: a group of those ISLs with a traffic load less than 25% of the total ISL capacity; a group of those ISLs with a traffic load between 25% and 50%; a group of those ISLs with a traffic load between 50% and 75%; and the group of ISLs with a traffic load more than 75%.

Traffic load distribution

Figure 8.3.14 shows that our GALPEDA distributed the traffic load in the LEO satellite network more evenly than GALP1. If we use GALPEDA to allocate the traffic, less than 10% of the ISLs have a traffic load of more than 75%. In contrast, if we use GALP1 about 20% of the ISLs have a traffic load of more than 75%. It means that GALPEDA provides a better traffic load distribution. In addition to that, when we use GALPEDA the number of lightly-loaded ISLs is higher than when we use GALP1. This improvement results in a higher reserved bandwidth for the future incoming traffic.

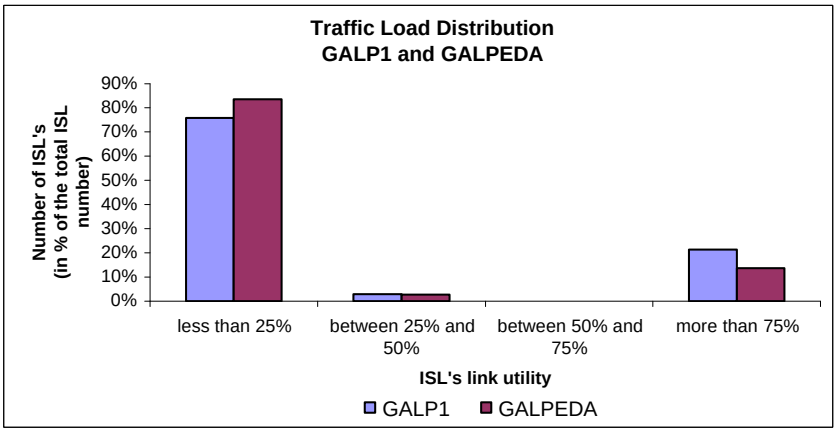

Figure 8.3.14: Traffic load distribution in GALP1 and GALPEDA

Arrival rate

If we increase the arrival rate from 10 packets per second to 50 packets per second, GALPEDA provide better traffic distribution than the GALP1 as shown in figure 8.3.15.

Figure 8.3.15 shows the traffic load distribution for GALP1 and GALPEDA. An increasing call arrival rate or number of calls results in a bigger performance difference between GALP1 and GALPEDA. If the call arrival rate is 50, then the number of ISLs with a traffic load of more than 90% is approximately 21% for GALPEDA, compared with approximately 35% for GALP1. This reduced the number of heavy loaded ISLs for about 14%. The number of ISLs with a traffic load of less than 50% is approximately 43% for GALP1 and 75% for GALPEDA. This increased the number of lightly-loaded ISLs for about 32%.

Multiclass traffic

In case of multiclass traffic, GALPEDA provides a shorter delay for high priority traffic; whereas in GALP1 there are no privileges given to high priority traffic. Figure 8.3.16 shows that GALPEDA provides a shorter path delay for high priority traffic than the average path length in GALP1. The average path length for low priority traffic in GALPEDA is longer than the average path length of GALP1.

Table 8.3.5 shows the minimum, average, and maximum value of the traffic path length, with

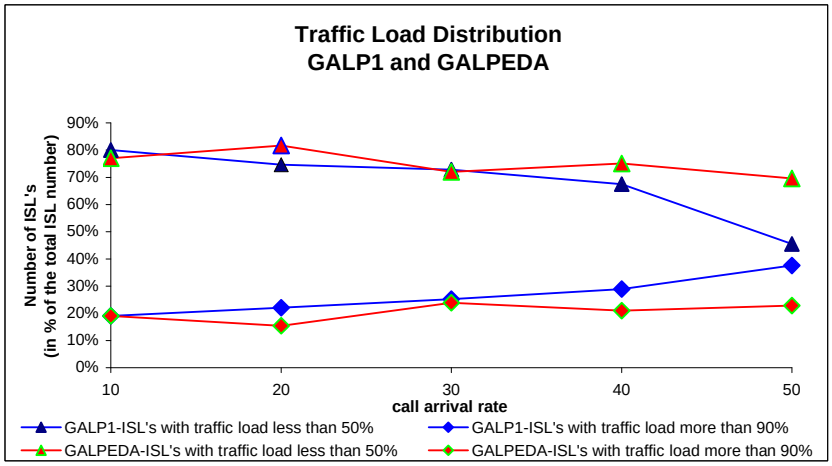

Figure 8.3.15: Traffic load distribution by using GALP1 and GALPEDA with the increased number of call arrival rate

various value of call arrival rate for both GALP1 and GALPEDA. GALP1 make no distinctions of traffic classes; we on the other hand, distinguish two traffic classes: low priority and high priority traffic.

Table 8.3.5: Average path length of different type of traffic for GALP1 and GALPEDA

arrival rate	GALP1 path length			GALPEDA					
				path length of Low priority traffic			path length of high priority traffic		
	min	average	max	min	avg	max	min	avg	max
10	2	3.17	5	2	2.67	4	2	2.63	5
20	2	3.2	5	2	3.36	5	2	2.66	4
30	2	3.23	5	2	3.4	5	2	2.67	5
40	2	3.46	5	2	3.68	5	2	3	5
50	2	3.57	5	2	3.67	5	2	3.4	5

As shown in this table (Table 8.3.5), GALPEDA provides on average a shorter path length for high priority traffic than GALP1 can provide. However, a longer path length can occur for low priority traffic if we use GALPEDA.

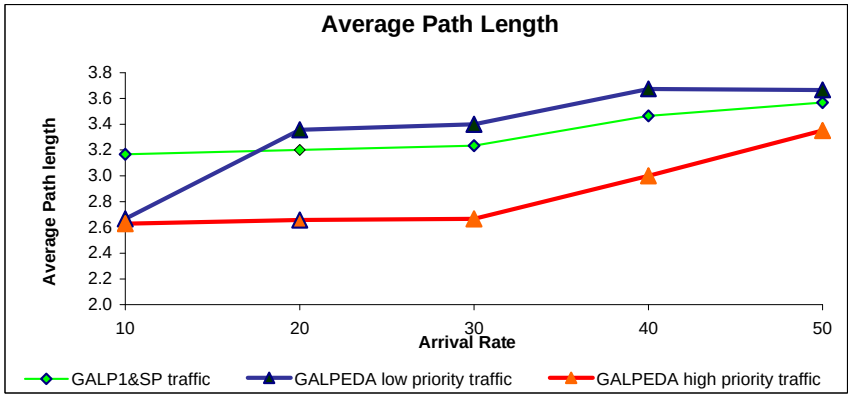

Figure 8.3.16: Average path length with the increase number of call arrivals for GALP1 and GALPEDA

Mutation and tabu tenure

In the previous case, the mutation probability is 10%. This means a mutation occurs once in every 10 periodical updatings. A mutation occurs by using the *previous solution without updating* as one of the parents.

We investigated the performance of different values of mutation probability as given in table 8.3.6. In this case, the simulation parameters remain the same for GALP1 and GALPEDA; but in GALPEDA, we vary the value of mutation probability from 10% to 30%. Since in GALP1 no mutation is possible, the values remain constant.

In GALPEDA, there is an improvement in the average path length of both low and high priority traffic when we increase the mutation probability from 10% to 15%. If we increase the mutation probability from 15% to 20%, the average path length in GALPEDA becomes longer but remains shorter than the traffic path length in GALP1. Once we used 25% as mutation probability, then the average length of low priority traffic becomes longer than the average path length in GALP1. The average path length for high priority traffic remains shorter than the average path length of GALP1. The average path length for high priority traffic in GALPEDA becomes longer than the average path length in GALP1 when we use a mutation probability of 30%. In this case, it seems that our mutation appears too frequently, that it provides a higher bias in the LP solver.

This effect is reduced by the implementation of our Tabu tenure procedure. In which a

saturated link will not be included in the alternative solutions for two iterations time. Even though the mutation procedure recommended these saturated links, the Tabu tenure does not allow these saturated links to become one of the alternative solutions of our GALPEDA. The Tabu tenure aborted the mutation procedure, and used only the current two chromosomes as the parents.

Table 8.3.6: Average traffic path length with various mutation probability

GALP1 traffic path length			mutation probability	GALPEDA					
				Low priority traffic path length			high priority traffic path length		
Min	avg	max		min	avg	max	min	avg	max
2	3.17	5	0.1	2	2.67	4	2	2.63	5
2	3.17	5	0.15	2	2.5	5	2	2.3	4
2	3.17	5	0.2	2	2.8	5	2	2.6	5
2	3.17	5	0.25	2	3.1	7	2	2.9	5
2	3.17	5	0.3	2	3.25	7	2	3	5

Processing time

In order to investigate the processing time required for GALPEDA, a comparison has been made between GALP1 and GALPEDA. The following table ((table 8.3.7)) shows the average processing time for the GALP1 algorithm and our GALPEDA algorithm. In this case we consider 16 satellites in 4 planes.

Table 8.3.7: Processing time of GALP and EDA with 16 satellites in 4 planes

	min (msec)	average (msec)	max (msec)
GALP1	35	49.41	60
GALPEDA	15	27.06	50

8.4 Discussion: Performance Analysis of GALPEDA

As shown in our simulation results, our proposed algorithm provided beneficial features on distributing and giving preferences for different traffic classes. A higher arrival rate increases the average path length difference between the two traffic classes. This is because we consider a constant threshold capacity of each ISL in our simulation. Similar occurrences is happened for the MMPP traffic model.

If we increased the value of λ^{low} from 1 to 20 as shown in figure 8.3.11 for the Poisson model, we achieved a higher average path length difference between low priority and high priority traffic between λ^{low} =1 to λ^{low} =10. Subsequently, no improvement is achieved when the value of λ^{low} is increased to 15; then, the difference of the average path length decreases if the value of λ^{low} is further increased to 20.

Apparently, if we increase λ^{low} to a value higher than the minimum value obtained from (7.2.13), a higher average path length difference can be achieved. However, a maximum value of $\lambda^{low}/\lambda^{high}$ needs to be considered. Starting from that value, the average path length difference decreases. We will define this maximum value of $\lambda^{low}/\lambda^{high}$ in the next paragraphs.

We consider a single link of capacity Q and assume that load on this link at time t is comprised of i calls. We consider a situation in which i calls are requesting service at time t with probability $p_I(t)$. $avg\ i$ denotes the average number of calls requesting service:

$$avg(i) = \sum_{i=1}^{\infty} p_i(t)i \tag{8.4.1}$$

Each call represents an application whose performance or utility π as a function of available bandwidth b, is given by the function $\pi(b)$. We assume that $\pi(0) = 0$. When an application receives zero bandwidth, the value becomes zero. We also assume that $\pi\ (\infty) = 1$, when an application receives as much bandwidth as it wants and the value becomes 1. In multiclass traffic, different applications have different levels of performance at bandwidth $0<b<1$; but in all cases $\pi\ (b)$ is a non-decreasing function. We model a call of high priority class was rejected as receiving zero bandwidth ($b = 0$) and so has zero utility ($\pi(0)$).

In the case that capacity Q is given for one class traffic, each call receives identical shared

bandwidth Q/i. If all calls are accepted, total utility of the satellite network is given by:

$$V(i) \equiv i\pi(\frac{Q}{i}) \tag{8.4.2}$$

In the case of the best effort architecture, if we use a greedy algorithm then the total utility of the system:

$$V_B(Q) = \sum_{i=1}^{\infty} p_i(t) i\pi(\frac{Q}{i}) \tag{8.4.3}$$

and, in the case of our proposed architecture, using a multiservice algorithm with reserved bandwidth i_{max} (Q). If i_{max} is reached then low priority calls requesting this link will be rejected. This i_{max} defines our maximum value of $\lambda^{low}/\lambda^{high}$. The total utility of the system is

$$V_R(Q) = \sum_{i=1}^{i_{\max}(Q)} p_i(t) i\pi(\frac{Q}{i}) + \sum_{i=i_{\max}(Q)}^{\infty} p_i(t) i_{\max}(Q)\pi(\frac{Q}{i_{\max}(Q)}) \tag{8.4.4}$$

When we use this inequality as below:

$$\frac{V_R(Q)}{i} \geqslant \frac{V_B(Q)}{i} \tag{8.4.5}$$

we will achieve the following

$$i_{\max}(Q)\pi(\frac{Q}{i_{\max}(Q)}) > (i_{\max}(Q)+1)\pi(\frac{Q}{(i_{\max}(Q)+1)}) \tag{8.4.6}$$

This inequality performs the constraints to our maximum value of $\lambda^{low}/\lambda^{high}$. Both inequalities, (7.2.13) and (8.4.6), perform the lower and upper bound of the value of $\lambda^{low}/\lambda^{high}$. In order to achieve a better performance of GALPEDA, we need to consider these constraints.

8.5 Summary

In this chapter we discussed the performance of our GALPEDA by comparing this algorithm with the previous work GALP1. We focused on the performance of our GALPEDA in distributing the traffic load and providing a privilege for high priority traffic for various types of traffic model and various satellite constellation parameters. The distribution of traffic load is

investigated by observing the number of ISLs with highly loaded traffic. If we used GALPEDA, the number of ISLs with highly loaded traffic is reduced by more than 25 % compared to the GALP1 algorithm. This resulted in more spare capacity in the other ISLs, which can be reserved for the next incoming high priority traffic. The privilege for high priority traffic is given by providing the high priority traffic with a shorter path than the low priority traffic. This resulted in a shorter delay time. If we used GALPEDA the average path length of high priority traffic is reduced by more than 25% compared with the average path length given to all types of traffic when GALP1 is used. This privilege for high priority traffic is very beneficial in the multiclass traffic allocation. GALPEDA distributed the traffic load more evenly and provided privilege for high priority traffic for both, Poisson and MMPP traffic model.

Chapter 9

CONCLUSIONS

9.1 Introduction

In this chapter, we summarize and discuss results that we achieved in our simulation. The simulation results, with various setting parameters of: GALPEDA, satellite constellation and traffic models, show the performance of our GALPEDA in allocation of multiclasses traffic in LEO satellite network. The achievements have been made by our novel traffic allocation algorithm, GALPEDA, in distributing traffic load more evenly will be presented, followed by recommendations for future that would build on the results of our thesis.

9.2 Summary

In this thesis, we consider LEO satellite constellation networks with ISLs. We focused on proposing a combination algorithm, GALPEDA, which allocates multiclass traffic in the ISLs of LEO satellite networks. The main objective of this GALPEDA is to give a privilege for the class of high priority traffic and to distribute the traffic load more evenly over the satellite network. The contributions of this thesis which have been made are:

1. We introduce two terms in our objective function: the first term minimizes delay due to the number of hops in a path, the second term maximizes the residual bandwidth over the whole network. The first term concerns more about the path propagation delays across simulated satellite networks. The second terms concerns more about the remaining

bandwidth of the ISLs in the simulated satellite networks. These two terms may sometimes conflict, since the minimum hop route will not necessarily minimizes the cost due to the bandwidth available. Therefore we introduce an original idea of a privilege parameter for each traffic class into this objective function. If we choose the value of this privilege parameter between its lower and upper bound values, this parameter will give a certain privilege for a class of high priority traffic.

2. In order to cope with the dynamic topology of LEO satellite network, we introduced a novel combination algorithm, GALPEDA. We divided the problem into two separate problems: periodical and incremental problem. In the first, we solve the problem of the whole satellite network. In the second, we solve the problem of a particular incoming traffic in a particular OD pair. Our novel combination algorithm solves these two problems subsequently. In periodical problem, the GALP-part of GALPEDA allocates the incoming traffic into ISLs of the whole satellite network. This tends to distribute traffic more evenly over the whole network. Furthermore, the EDA-part of GALPEDA allocates incoming traffic into ISLs between origin satellite and destination satellite. Another original idea is the use of mutation and tabu tenure property in order to improve the performance of our GALPEDA. This mutation property is used in order to escape from the local optima. With a certain mutation probability, we use an unmodified solution of the previous periodical updates as one of the parents in the cross over procedure. Since the satellites move into their new positions, this mutant parent introduces a new type of chromosome in the cross over. In case that the mutation probability is high, the mutation becomes too often. This results in a poor performance of our GALPEDA. Tabu tenure property of tabu search helps us to reduce this negative impact of mutation procedure. If an ISL has been considered to be tabu, this ISL cannot be used for the period of its tabu tenure. In case this ISL becomes an alternative solution suggested by mutation procedure, tabu tenure property will prohibit the use of this ISL. This results in controlling the effect of a high mutation probability.

3. The predictable satellite positions in LEO satellite network can be used to improve the handover procedure. The LEO satellite topology is updated in the periodical updating

time. At this time, we calculate the new positions of the satellites. The solutions to the traffic allocation problem from the previous time interval are evaluated. We hand over the previous solutions in new solutions, by considering the new positions of satellites. In case a solution for an OD satellite pair is required to be handed over, we keep the original OD path and add two extra ISLs. The first ISL connects the old origin satellite and a new origin satellite. The second ISL connects the old destination satellite and a new destination satellite. The resulting solution becomes the first solution in the initial population of GALPEDA. Instead of starting with an empty initial population, the GALPEDA starts with a modified version of the previous time intervals solution. This original idea of introducing the chromosome of the previous solution into the initial population of GALPEDA, accelerates the processing time of GALPEDA; since the solution from the previous time interval provides information about the latest condition of the whole satellite network.

4. Our GALPEDA is used for two types of traffic model, Poisson and MMPP. Our GALPEDA can cope with the bursty property of the nowadays traffic condition. In addition to that we introduce an original adaptive length of time interval. The length of a time interval depends on the traffic load at this time interval. If the traffic load is high then the time interval becomes shorter, or the periodic updates will be done more frequently. Otherwise the time interval becomes longer. This property can reduce the overhead cost of periodical updating.

In this thesis, we have developed and analyzed a novel method of allocating multiclass traffic in a dynamic LEO satellite network. We performed a simulation with various setting parameters to evaluate the performance of our GALPEDA.

Firstly, we considered various setting of GALPEDA parameters: population size, hop limit, and node degree. If we used a smaller population size, the algorithm converged relative faster than a higher population size. But, on the other hand a higher population size gives a better value of end solution than a smaller population size. Increased size of the population means that there is more combination and choices of chromosomes to perform a cross over. Therefore, we need a longer time to approach the equilibrium but the result is more accurate. In order to distribute traffic load more evenly, a higher population size can be used. In addition, if we

consider different values of the hop limit, we see that a smaller hop limit has a longer progress speed than a higher hop limit. On the other hand, we need to consider different values of the hop limit for different types of traffic. The difference between these hop limits defines the spare bandwidth for different traffic classes. Therefore, the minimum and maximum value of λ^{low} $/\lambda^{high}$ has been considered.

Secondly, we considered various settings of satellite parameters: number of satellites and number of planes. By increasing the number of satellites, our GALPEDA distributes traffic more evenly and discriminates more between different traffic classes. There is a tendency to achieve a better value of end solution by decreasing the number of planes in satellite network. Complexity in the inter-plane handover procedure contributes into this drift. Due to the movement of satellites in neighboring planes, ISLs are switched on and off between satellites in the adjacency plane. Therefore, an additional handover procedure needs to be performed, which delivers a higher error rate in the system. Alternatively, increasing the number of planes will not affect path length and distribution of traffic, but it effects the processing time and the accuracy of the results. It takes longer to calculate the path using links in different planes rather than using links in the same plane.

9.3 Future Work

In this thesis we concentrated more into the performance of our GALPEDA in order to allocate multiclass traffic in LEO satellite constellation. Future work based on this thesis can be done in the areas as follows:

In our model, we have not compared the performance of using three different methodologies for allocating traffic in a satellite constellation. The first method is to find the best path for an incoming packet. When a connection is requested, the origin's satellite searches an available route for this request in the available path directories. If there is still an available transmission link to accommodate this request then use this path. If it is not the case then allocating procedure is initiated. This method provides fast processing time. However, if congestion occurs in other parts of the world, a congested link is not available (for the same request) till a periodically updating is initiated. Another method is to find the best path for the whole

network when a call is coming. When there is an incoming request, like above, the origin's satellite searches for an available route for this request in the current path directories. If there is no available transmission link in the database to satisfy this demand, then a routing procedure is initiated. Unlike in the first method, not only the best path from the origin to the destination is searched, and also the optimal alternative route for the whole network is searched. The benefit is that when there is congestion in other parts of the world the procedure will not wait till the periodically updating and the routing procedure will optimize the whole network. The drawback of this method is that the processing time is slower than in the previous one. The last method is to find the best path for the whole network when one transmission link is saturated. In this method, routing procedure is started directly when there is a saturated transmission link. Based on this method, it solves instantly when there is a saturated transmission link. In case another request comes into network, the network is ready to allocate this request. Hopefully the advantages of all these methods can be achieved, and disadvantages of these methods can be eliminated.

A research in the direction of a subspace search can be performed in the future. In the subspace search we could use the zones in our LEO satellite footprints. The use of subspace can result in a faster processing time. On the other hand, a more complex processing need to be performed on the boundary satellites. A neighborhood search can be used to solve the whole network problem. In addition to this, if we perform subspace search, there is a possibility to divide the zones into a daylight and nighttime zone. This separation can be used to define the privilege parameter value for each of the zones.

In our research, we have not considered self similar traffic model. We assumed that the current traffic can be modeled as MMPP traffic model. Even though MMPP traffic model can be used to model a current traffic, a study in this direction can be useful. Furthermore, in our research we considered only two traffic classes. Consequently, there is a need to model traffic in more than two traffic classes as proposed in the IPv6 traffic classes.

Since a satellite communication is used as a part of the communication system, which we can integrate our model with terrestrial network and perform a hybrid communication system, a more beneficial model can be achieved.

Appendix A

QUEUEING MODELS

A.1 Queueing models

Blocking methods are given in [131]. The first model is a Generalized Engset Loss Station. We introduce a set, C, of call classes (for example: voice class, short traffic class and long traffic class). The arrival pattern of class c calls is a Poisson process with rate λ_c. We denote by $n_c(t)$ the number of class c calls in progress at time t.. Let L denotes a set of server types (or channel types). There are $S_l$channels of type $l \in L$. A class c call requires holding A_{lc} servers of type l simultaneously. Call holding time distribution is a general distribution $G_c(t)$ with mean $1/\mu_{\ c.}$

The second model is a Queuing Loss Station. This station combines the Loss station and the Queuing station. This model of queuing loss station is used in this research. This model is only for one satellite. The whole satellite network is composed of a number of these Queuing Loss stations.

In Erlang Loss Model, as given in [21] the stationary distribution of the number of busy servers is given by:

$$P(n) = \frac{1}{G(s)} \frac{a^n}{n!}, 0 \leq n \leq S \tag{A.1.1}$$

where

$$a = \frac{\lambda}{\mu} \tag{A.1.2}$$

is the offered load and

$$G(S) = \sum_{n=0}^{S} \frac{a^n}{n!} \tag{A.1.3}$$

, is normalization constant.

A.2 Congestion

Use of Erlang Loss Formula can give the probability that all servers are found busy in the steady state, it is called in [131] to be *Time Congestion* (since this represents the proportion of time that all servers are busy):

$$B(S) = P(S) = \frac{a^S}{S!} \left[\sum_{i=0}^{S} \frac{a^i}{i!} \right]^{-1} = 1 - \frac{G(S-1)}{G(S)} \tag{A.2.1}$$

The probability that a newly arriving call finds all servers occupied and hence is lost or blocked, i.e. leaves the system without being served, is called in [131] to be *Call Congestion* or *call Loss Probability*. In the Erlang Model, because the arrival process is Poisson then the call congestion and time congestion will be equivalent [131].

The Time Congestion $B(S, N)$ for a class c is given by:

$$B(S, N) = P(S, N) = 1 - \frac{G(S-1, N)}{G(S, N)} \tag{A.2.2}$$

The Call Congestion $L(S, N)$ is

$$L(S, N) = B(S, N-1) = 1 - \frac{G(S-1, N-1)}{G(S, N-1)} \tag{A.2.3}$$

In a generalized Engset Loss Model, arriving calls do not form an infinite Poisson process (with rate λ and mean $1/\mu$. We replace the Poisson arrival by a finite number N of sources $(N > S)$).

In the Erlang Loss model, the stationary distribution of the number of busy lines is given in [131] as

$$P(n) = \frac{1}{G(S)} \frac{a^n}{n!}, 0 \leq n \leq S \tag{A.2.4}$$

Where $a = \lambda/\mu$ and $G(S)$ is a normalization constant given by

$$G(S) = \sum_{n=0}^{S} \frac{a^n}{n!} \tag{A.2.5}$$

Time congestion, which represents the proportion of time that all lines are busy, is

$$L_{time} = P(S) = 1 - \frac{G(S-1)}{G(S)} \tag{A.2.6}$$

Call congestion or call loss probability is defined as the probability that a newly arriving call finds all lines is occupied; hence this call is blocked. Call congestion of the Erlang Loss Model is similar to time congestion following the Poisson Arrivals See Time Averages (PASTA) property [164]. In the Generalized Engsett Model which is used in our model, we replace Poisson arrival by a finite number N of sources ($N>S$). Each source generates a call with

$$P(\overline{n} \mid S, N) = \frac{1}{G(S,N)} \prod_{c \in C} \left[\begin{array}{c} N_c \\ n_c \end{array} \right] b_c^{n_c}, n \in F(S,N) \tag{A.2.7}$$

an exponentially distributed intergeneration time with mean $1/\mu$.

In this model, $n(t)$, the number of calls in progress at time t, has in steady state the following distribution [131], which $b = \nu/\mu$ and $F(S,N)$ is the set of feasible states given by

$$F(S,N) = \overline{n} \geq 0 : \sum A_c n_c(t) \leq S, n_c \leq N_c, c \in C \tag{A.2.8}$$

and normalization constant given by

$$G(S,N) = \sum_{n \in F(S,N)} \prod_{c \in C} \left[\begin{array}{c} a_c^{n_c} \\ n_c! \end{array} \right] \tag{A.2.9}$$

Time Congestion L_{time} (S,N) for a class c is given by:

$$L_{time}(S,N)_c = P_c(S,N) = 1 - \frac{G(S-1,N)}{G(S,N)} \tag{A.2.10}$$

and Call loss probability is given by

$$L_{call}(S,N)_c = P_c(S,N) = 1 - \frac{G(S-1,N-1)}{G(S,N-1)} \tag{A.2.11}$$

Based on these probabilities of incoming calls for each class it is possible to look at the performance of our multiservice routing algorithm.

Appendix B

TRAFFIC MODEL

B.1 Traffic Models

In order to model the todays communication traffic, we describe in this appendix various traffic processes. There are two types of traffic processes as given by [12]:

1. Point processes consist of the sequence of arrival instants. Bearing in mind the number of units in one arrival there should be two types of point processes:

   - Simple Traffic consists of single arrivals of discrete entities (packets, cells, frames, etc.), and the traffic can be mathematically described as a point process, which contains a sequence of arrivals measured from the origin t0. There are two equivalent descriptions of point processes, namely:

     (a) A counting processes: this is a continuous-time, non-negative integer-valued stochastic process, where $N(t)$ is the number of traffic arrivals in the interval $(0, t]$.

     (b) An interarrival time process is a real valued random sequence $\{A_n\}$ where $A_n = T_n - T_n - 1$ is the length of the time interval separating the nth arrival from the previous one.

   - Compound traffic consists of batch arrivals; that is the arrivals may consist of more than one unit at an arrival instant T_n. Also, it is sometimes useful to incorporate the

notion of workload into the traffic description. The workload is a general concept describing the amount of work W_n brought to a system by the nth arriving unit, which is usually assumed independent of interarrival times and batch sizes. An example is compressed video, is also known as coded video. Video transmission is normally transmitted by considering the redundancy in the digitized pictures, by compressing each frame into a fraction of its original size. The compressed frames have random sizes (bit rates), which are then transported over the network and decoded, at their destination. This kind of video traffic is referred to as VBR (Variable Bit Rate) video, which must be delivered every 1/30 of a second. Then the workload consists of the coded frame sizes, since a frame is roughly proportional to its transmission time.

2. Stochastic intensity processes, primarily of theoretical interest, describes the random rate at which point arrivals occur in N, given the past history of N and possibly additional information.

B.2 Point Processes

In our research, only point processes will be considered. Based on this point process some traffic models have been designed. The models vary from models that are based on the renewal traffic processes to models called fluid traffic models that are based on the compound traffic.

Arvidsson et al [11] and Jagerman et al. [12] have been studied various traffic models as given below:

Renewal Traffic Models: This model is a point process in which the interarrival time An is IID (independent, identically distributed) and their distribution can be general. Renewal processes are relatively simple because they are independent of each other, and then the correlation between the arrivals becomes zero. This correlation carries information about the 'burstiness' of the traffic. Models which have auto correlated nature of traffic are essential for predicting the performance of emerging broadband networks. There are two types of models in these renewal traffic models - Poisson models and discrete time analogue Bernoulli models:

Poisson models: the oldest traffic models are the Poisson models, which can be characterized

as renewal processes whose interarrival times are exponentially distributed, with rate parameter λ that is

$$P\{A_n < t\} = 1 - \exp(-\lambda_t) \tag{B.2.1}$$

Poisson processes have some useful properties: the superposition of independent Poisson processes results in a new Poisson process whose rate is the sum of the component rates. And, the independent increment property gives the memory-less properties of Poisson processes. It is possible to define a time dependent Poisson process by letting the rate parameter λ depend on time. One variant of the Poisson process, which tries to contribute the correlation between interarrival times from two underlying states, is SPP (Switched Poisson Process). In this process the rate is determined by the state of an underlying two state Markov chain. The transition rates of the Markov chain, denoted by rate1 (from state 1 to 2) and rate2 (from state 2 to 1), while the arrival rates associated to the two states are independent as λ_1 and λ_2 respectively. Switched Bernoulli Process is a discrete version of SPP.

Bernoulli models: this is the discrete-time analogue of the Poisson processes. Here the probability of an arrival in any time slot is p, is independent of any other one. Some versions of this Poisson and Bernoulli process have been studied in the previous papers [12], namely:

Markov Based Traffic Models: As already mentioned, the renewal traffic model has the drawback that it cannot be used to transfer the information of the 'burstiness' of the traffic. Therefore, Markov Based Traffic Models will be used to introduce the dependence in the random sequence $\{An\}$, which can potentially capture 'burstiness' information. Suppose,

$$M = \{M(t)\}_{t=0}^{\infty} \tag{B.2.2}$$

is a Markov process with a discrete state space, then the rate parameter depends on the state from where the jump occurs. It follows the transition matrix $P = [pi, j]$. There are some variants of these models such as [165]:

Autoregressive Markov Model (AMM): generates a Markov dependent number of arrivals within a frame. It defines the next variety in the sequence as an explicit function of the previous variants within the time window. Assume the number of arrivals in frame n by λ_n then $\lambda_n = a\lambda_{n-1} + bw(n)$, where a and b are constants, and w is a sequence of independent normally

distributed random variables with mean η and variance σ^2. It is called the linear autoregressive model. Another autoregression model can be used such as Moving average models and ARMA models.

Markov Hyperexponential Model [166,167], in which a suggestion is made for 'bursty' traffic. It is assumed that the burst lasts for a negative exponentially distributed time with mean $1/\mu$, while silence periods have two hyper exponential distributions with parameters λ_1, λ_2 and α. The two distributions represent two different types of silence, pauses between talks and pauses when other party of a conversation is being active.

Markov Modulated Processes: the idea is to introduce an explicit notion of state into the description of a traffic stream. Its current state will control or modulate the probability of the traffic mechanism. While M is in state k then the probability of traffic arrivals is completely determined by k, and this holds for every $1 < k < m$. This model is less analytical, tractable and is more complex in the modeling process.

Markov Modulated Poisson Process (MMPP): This is almost the same as SPP (the difference is that in SPP the transitions rate is also dependent of the state transitions), combining the simplicity of the modulating of Poisson Process into the Markov Process. In this model each state k of M, arrivals occur according to a Poisson process at rate λ_k, when the state changes so does the rate.

Transition Modulated Processes: the modulating agent in this process is the state transition rather than a state. State transition occurs in the state boundaries and is modulated by mxm Markov transition matrix $P = [p_{i,j}]$. This is also similar to SPP, in which the arrival rate is constant.

Fluid Traffic Models: this model, instead of traffic units, it views traffic as a stream of fluid, which is characterized by flow rate (bits per second), so that the traffic unit becomes traffic volume. This is appropriate for the cases where individual units are Big in numbers, relative to the chosen time scale. For example, in the video transmission, a compressed high quality video frame could consist of a thousand cells, so that the impact of an individual cell is negligible. When the traffic change has a slow rate, using this model can save a lot of CPU time and memory, because the traffic can be modeled as Markov modulated model with constant rate traffic.

Self Similar Traffic Models: this model based on packet traffic, which is characterized by the statistically self similar or fractal in nature. In this traffic model, the traffic looks statistically almost similar over a wide range of time intervals.

Bibliography

[1] In-Stat/MDR, Event Horizon: Two Billion Mobile Subscribers by 2007. 2003 Subscriber Forecast, 2003, available: www.instat.com/rh/wirelessweek.

[2] In-Stat/MDR, Mobile Messaging in Japan, Asia and AME: 2003 Through 2007, 2003, available: www.instat.com.

[3] G. McMahon, R. Septiawan and S. Sugden, "Class Dependent Traffic Allocation in a LEO Satellite Network." Telecommunication Systems 22(1): 2003, pp. 241-266.

[4] T. Kwon, Y. Choi, C. Bisdikian, et al., "Call Admission Control for Adaptive Multimedia in Wireless/ Mobile Networks." WOWMOM 98, Dallas, Texas, USA, ACM, 1998:pp 111-116.

[5] B. Vandalore, R. Jain, S. Fahmy, et al., "AQuaFWiN: Adaptive QoS Framework For Multimedia in Wireless Networks and Its Comparison with Other QoS Frameworks." LCN '99, 1999.

[6] G. Frankhauser, M. Dasen, N. Weiler, et al., "WaveVideo – An Integrated Approach to Adaptive Wireless Video." ACM Journal on Mobile Networks and Applications 4(4): 1999, pp. 255-271.

[7] A. Jamalipour, The Wireless Mobile Internet: Architectures, Protocols and Services. West Sussex, England: John Wiley&Sons Ltd. 2003.

[8] G. Apostolopoulos, R. Guerin, S. Kamat, et al., "On Reducing the processing cost of on demand QoS path computation." Journal of High Speed Networks 7(2): 1998, pp. 77-98.

[9] R. A. Guerin and A. Orda, "QoS Routing in Networks with Inaccurate Information: Theory and Algorithms." IEEE/ACM Transactions on Networking 7(3): 1999, pp. 350-364.

[10] G. McMahon, R. Septiawan and S. Sugden, "A Multi-service Traffic Allocation Model for LEO Satellite Communication Networks." IEEE Journal on Selected Areas in Communications: to be published in 2004.

[11] A. Arvidsson and R. Harris, "Analysis of the Accuracy of Bursty Traffic Models." First International Conference on Telecommunication System Modeling and Analysis, Nashville, Tenesse, USA, Centre for telecommunication network research, 1993:pp 206-211.

[12] D. L. Jagerman, B. Melamed and W. Willinger, Stochastic Modelling of Traffic Processes. in Frontiers in Queuing: Models, Methods and Problems. J. Dsalalow, CRC Press: 271-320, 1997.

[13] T. Janevski, Traffic Analysis and Design of Wireless IP Networks. Norwood, MA: Artech House, Inc 2003.

[14] J. J. Bae, T. Suda and R. Simha, "Analysis of a finite buffer queue with heterogeneous markov modulated arrival process: A study of the effects of traffic burstiness on individual packet loss." IEEE INFOCOM, IEEE INFOCOM, 1992:pp 219-230.

[15] A. T. Andersen and B. F. Nielsen, "A Markovian approach for modeling packet traffic with long-range dependence." IEEE Journal on Selected Areas in Communications 16(5): 1998, pp. 719-732.

[16] K. K. Leung, W. A. Massey and W. Whitt, "Traffic Models for Wireless Communication Networks." IEEE Journal on Selected Areas in Communications 12(8): 1994, pp. 1353-1364.

[17] D. Lam, D. Cox and J. Widom, "Teletraffic Modeling for Personal Communications Services," IEEE Communications Magazine: Special Issues on Teletraffic Modeling Engineering and Management in Wireless and Broadband Networks, 35: pp. 79-87, February 1997 1997.

[18] M. Roughan, A. Greenberg, C. Kalmanek, et al., "Experience in Measuring Backbone Traffic Variability: Models, Metrics, Measurements and Meaning." ACM SIGCOMM Internet Measurement Workshop, 2002:pp 91-92.

[19] A. Klemm, C. Lindemann and M. Lohmann, "Traffic Modeling and Characterization for UMTS Networks." Globecom, Internet Performance Symposium, San Antonio, TX, USA, 2001:pp 1741-1746.

[20] S. Kasahara, "Internet Traffic Modelling: Markovian Approach to Self-Similar Traffic and Prediction of Loss Probability for Finite Queues." IEEE Trans. On Communications E84-B(8): 2001, pp. 2134-2141.

[21] H. Kobayashi, S.-Z. Yu and B. L. Mark, "An Integrated Mobility and Traffic Model for Resource Allocation in Wireless Networks." WoWMoM 2000, Boston, MA, USA, ACM, 2000:pp 39-47.

[22] H. Arsham, System Simulation: The shortest distance from learning to applications, available: http://ubmail.ubalt.edu\~harsham/simulation/sim.htm.

[23] L. S. Golding, "Satellite Communications Systems Move into The Twenty-First Century." Wireless Networks 4: 1998, pp. 101-107.

[24] T. R. Henderson and R. H. Katz, "Network Simulation for LEO Satellite Networks." American Institute of Aeronautics and Astronautics (AIAA) International Communications Satellite Systems Conference (ICSSC), 2000, AIAA, 2000.

[25] I. Ali, N. AL-Dhahir and J. E. Hershey, "Doppler Characterization for LEO satellites." IEEE Transactions on communications 46(3): 1998, pp. 309-313.

[26] B. H. Fleury, M. Tschdin, R. Heddergott, et al., "Channel Parameter Estimation in Mobile Radio Environments Using The SAGE Algorithm." IEEE Journal on Selected Areas in Communications 17(3): 1999.

[27] Z. Sun, T. Ors and B. G. Evans, "ATM-over Satellite Demonstration of Broadband Network Interconnection." Computer Communications 21: 1998, pp. 1090-1101.

[28] A. Ganz, Y. Gong and B. Li, "Performance Study of Low Earth Orbit Satellite Systems." IEEE Transactions on Communications 42(2/3/4): 1994, pp. 1866-1871.

[29] B. Gavish and J. Kalvenes, "The Impact of Satellite Altitude on The Performance of LEOS Based Communication Systems." Wireless Networks 4(2): 1998, pp. 199-212.

[30] R. Bekkers and J. Smits, Mobile Telecommunications: Standards, Regulation, and Applications: Artech House Boston-London 1997.

[31] M. Werner, A. Jahn, E. Lutz, et al., "Analysis of System Parameters for LEO/ICO-Satellite Communication Networks." IEEE Journal on Selected Areas in Communications 13(2): 1995, pp. 371-381.

[32] L. Wood, G. Pavlou and E. B., "Effects on TCP Routing Strategies in Satellite Constellations," IEEE Communications Magazine, March 2001 2001.

[33] L. Breslau and S. Shenker, "Best Effort versus reservations: a simple comparative analysis." SIGCOMM '98, Vancouver, Canada, 1998.

[34] I. C. Paschalidis and J. N. Tsitsiklis, "Congestion Dependent Pricing of Network Services." IEEE/ACM Transactions on networking 8(2): 2000, pp. 171-183.

[35] S.-J. Yang, "Performance Evaluation of Routing Algorithms under Various Network Configuration Parameters." International Journal of Network Management 7: 1997, pp. 183-197.

[36] C. Pornavalai, G. Chakraborty and N. Shiratori, "QoS Based Routing Algorithm In Integrated Services Packet Networks." Journal of high speed networks 7: 1998, pp. 99-112.

[37] E. Modiano, "Random Algorithms for Scheduling Multicast Traffic in WDM Broadcast and Select Networks." IEEE/ACM Transactions on Networking 7(3): 1999, pp. 425-434.

[38] M. Alanyali and E. Ayanoglu, "Provisioning algorithms for WDM optical networks." IEEE/ACM Transactions on networking 7(5): 1999, pp. 767-778.

[39] S. Seetharaman, A. Durresi and R. Jain, "Signaling Protocol for Lightpath Provisioning." 26th Annual IEEE Conference on Local Computer Networks, 2001 (LCN 2001), Ohio, OhioState University and Nayna Networks, 2001:pp 82-88.

[40] A. Bremler-Barr, Y. Afek and S. Har-Peled, "Routing with a clue." SIGCOMM '99, Cambridge, MA, USA, 1999.

[41] A. Orda, "Routing with End-to-End QoS Guarantees in Broadband Networks." IEEE/ACM Transactions on networking 7(3): 1999, pp. 365-374.

[42] A. Shaikh, J. Rexford and K. G. Shin, "Load Sensitive Routing of Long Lived IP Flows." SIGCOMM '99, Cambridge, MA USA, 1999.

[43] G.-C. Lai and R.-S. Chang, "Support QoS in IP over ATM." Computer Communications 22: 1999, pp. 411-418.

[44] I. Chlamtac and A. Farago, "A new approach to the design and analysis of peer to peer mobile networks." Wireless networks 5: 1999, pp. 149-156.

[45] B. Sarikaya and M. Ulema, "An Evaluation of Quality of Service Characteristics of PACS Packet Channel." Mobile Networks and Applications 4: 1999, pp. 289-300.

[46] S. Lu, V. Bharghavan and R. Srikant, "Fair Scheduling in Wireless Packet Networks." SIGCOMM '97, Cannes, France, 1997:pp 473-489.

[47] J. A. Cobb, M. G. Gouda and A. El-Nahas, "Time Shift scheduling- Fair Scheduling of flows in high speed networks." IEEE/ACM Transactions on networking 6(3): 1998, pp. 274-285.

[48] S. Iatrou and I. Stavrakakis, "A dynamic Regulation and Scheduling Scheme for Real Time Traffic Management." IEEE/ACM Transactions on Networking 8(1): 2000, pp. 60-70.

[49] G. R. Ash and B. D. Huang, "An analytical model for adaptive routing networks." IEEE Transactions on communications 41(11): 1993, pp. 1748-1759.

[50] K. Shiomoto and N. Yamanaka, "Dynamic Burst Transfer Time-Slot-Base Network," IEEE Communications Magazine: pp. 88-96, October 1999.

[51] M. B. Pursley, H. B. Russell and P. E. Staples, "Routing for Multimedia Traffic In Wireless Frequency Hop Communication Networks." IEEE Journal on Selected Areas in Communications 17(5): 1999, pp. 782-793.

[52] H. Cruickshank, Z. Sun, L. Wood, et al., Report on LEO Satellite Network (Esprit Project EP28425, BISANTE: Broadband Integrated Satellite Network Traffic Evaluations), 4 August 1999,

[53] E. Lutz, "Issues in Satellite Personal Communication Systems." Wireless networks 4: 1998, pp. 109-124.

[54] A. Hung, M. J. Montpetit and G. Kesidis, "ATM via Satellite: A Framework And Implementation." Wireless networks 4: 1998, pp. 141-153.

[55] A. Ganz and Y. Gao, "Efficient Algorithms for SS/TDMA Scheduling." IEEE Transactions on Communications 40(8): 1992, pp. 1367-1374.

[56] A. Ganz and Y. Gao, "SS/TDMA Scheduling for Satellite Clusters." IEEE Transactions on Communications 40(3): 1992, pp. 597-603.

[57] L. Wood, G. Pavlou and B. Evans, "Managing Diversity with Handover To Provide Classes of Service in Satellite Constellation Networks." The AIAA International Communcaiton Satellite System Conference (ICSSC'01), Toulouse France, Centre for Communication System Research, University of Surrey, 2001.

[58] I. F. Akyildiz, H. Uzunalioglu and M. D. Bender, "Handover management in Low Earth Orbit (LEO) satellite networks." Mobile Networks and applications 4: 1999, pp. 301-310.

[59] J. D. Bakker and R. Prasad, "Handover in a virtual cellular network." IEEE Vehicular Technology Conference (VTC) 1999-Fall, Amsterdam, The Netherlands, Institute of electronic system, Aalborg University Denmark, 1999.

[60] E. D. Re, R. Fantacci and G. Giambene, "Efficient Dynamic Channel Allocation Techniques with Handover Queueing for Mobile Satellite Networks." IEEE Journal on Selected Areas in Communications 13(2): 1995, pp. 399-407.

[61] E. d. Re, R. Fantacci and G. Giambene, "Handover Queueing Strategies with Dynamic and Fixed Channel Allocation Techniques in Low Earth Orbit Mobile Satellite Systems." IEEE Transactions on Communications 47(1): 1999, pp. 89-102.

[62] H. Uzunalioglu, I. F. Akyilidiz, Y. Yesha, et al., "Footprint Handover Rerouting Protocol For Low Earth Orbit Satellite Networks." Wireless Networks 5: 1999, pp. 327-337.

[63] M. Emmelmann, H. Brandt, H. Bischl, et al., "An Access Protocol for Mobile Satellite Users with Reduced Link Margins and Contention Propability (invited paper)." First International Conference on Advanced Satellite Mobile Systems (ESA ASMS '03), ESA ESRIN, Frascati, Italy, 2003.

[64] M. Emmelmann and H. Bischl, "An Adaptive MAC Layer Protocol for ATM-based LEO Satellite Networks, Invited Paper." Vehicular Technology Conference (VTC 03), Orlando, Florida, USA, 2003.

[65] A. Ibrahim and S. Tohme, "A modified CDMA/PRMA Medium Access Control Protocol for Integrated Services in LEO Satellite Systems." WoWMoM 2000, Boston, MA USA, ACM, 2000:pp 93-100.

[66] C. Ward, C. H. Choi and T. F. Hain, "A Data Link Control Protocol for LEO Satellite Networks Providing A Reliable Datagram Service." IEEE/ACM Transactions on Networking 3(1): 1995, pp. 91-103.

[67] H. S. Chang, B. W. Kim, C. G. Lee, et al., "FSA-Based Link Assignment and Routing in Low-Earth Orbit Satellite Networks." IEEE Transactions on Vehicular Technology 47(3): 1998.

[68] C. Chen, E. Ekici and I. F. Akyildiz, "Satellite grouping and routing protocol for LEO/MEO satellite IP networks." WoWMoM'02, Atlanta, Georgia, USA, ACM, 2002.

[69] E. Ekici, I. F. Akyildiz and M. D. Bender, "A Distributed Routing Algorithm for Datagram Traffic in LEO Satellite Network." IEEE/ACM Transactions on Networking 9(2): 2001, pp. 137-147.

[70] O. Ercetin, S. Krishnamurthy, S. Dao, et al., "A Predictive QoS Routing Scheme for Broadband Low Earth Orbit Satellite Networks." Personal, Indoor and Mobile Radio Communications, London, UK, 2000.

[71] O. Ercetin, S. Krishnamurthy, S. Dao, et al., "Provision of Guaranteed Services in Broadband LEO Satellite Networks." Computer Networks 39: 2002, pp. 61-77.

[72] L. Wood, A. Clerget, I. Andrikopoulos, et al., "IP Routing Issues in Satellite Constellation Networks." International Journal of Satellite Communications Special Issue on Internet Protocols over Satellite 18(6): 2000.

[73] I. F. Akyildiz, E. Ekici and M. D. Bender, "MLSR: A Novel Routing Algorithm for Multi-Layered Satellite IP Networks." IEEE/ACM Transactions on Networking 10(3): 2002, pp. 411-424.

[74] F. Filali, G. Aniba and W. Dabbous, "Efficient Support of IP Multicast in the Next-Generation of GEO satellites." IEEE Journal on selected areas in communications special Issues on Broadband IP network via Satellite: to be pubished in 2004.

[75] J. Janssen, D. D. Vleeschauwer, G. H. Petit, et al., "Delay Bounds for Voice over IP Calls Transported over Satellite Access Networks." Mobile Networks and applications 7: 2002, pp. 79-89.

[76] E. Ekici, I. F. Akyildiz and M. D. Bender, "Network Layer Integration of Terrestrial and Satellite IP Networks over BGP-S." Proceedings of GLOBECOM 2001, 2001.

[77] H. Brandt, F. Krepel and C. Tittel, "Multiple Access Layer and Signalling Simulator for a LEO Satellite System." AIAA International Communication Satellite System Conference and Exhibition, Montreal, Canada., 2002.

[78] M. El-Kadi, S. Olariu and P. Todorova, "Predictive Resource Allocation in Multimedia Networks." IEEE Globecom, San Antonio, Texas, IEEE, 2001.

[79] M. Weiser, "The Computer for the 21st century." Scientific American: Mobile Computing and Communications Review 265(3): 1991, pp. 66-75.

[80] A. Cortese, Analysis: Third generation Wireless Promises High-Speed Alternative, available: http://www.cnn.com/2000/TECH/computing/10/13/3g.wireless.dreams.idg.index.html.

[81] J. Korhonen, Introduction to 3G Mobile Communications. Norwood, MA: Artech House, Inc 2003.

[82] W. Webb, Understanding Cellular Radio: Artech House Boston-London 1998.

[83] W. Stallings, Wireless Communications and Networks. Upper Saddle River, New Jersey: Prentice Hall 2002.

[84] W. X. Wang and S. D. Blostein, "Video Image Transmission Over Mobile Satellite Channels." Elsevier Signal Processing Image Communication 16: 2001, pp. 531-540.

[85] H. ITU-T Recommendation, Video Codec for Audiovisual Services at p* 64kbit/s, available: http://www.iso.ch/iso.

[86] M. Liou, "Overview of The px64 kbit/s Video Coding Standard." Communications of ACM 34(4): 1991, pp. 59-63.

[87] ISO/IEC13818-2, Generic Coding of Moving Pictures and Associated Audio, Part2: Video. ITU-T Recommendation H.262, 1998.

[88] H. ITU-T RecommendationH.263, "Video Coding for Low Bit Rate Communications." 1998.

[89] M. Z. Zhao and Y. C. Loh, "Multimedia Conferencing Designed with H.32x Capabilities." IEEE ICON 1999, Australia, 1999.

[90] ISO/IECIS13236, ITU-T Information Technology-Quality of Service: Framework. ITU-T Recommendation X.641, 1997.

[91] O. Espvik, Franken, T. Jansen, et al., "An Euroscom QoS Framework for Multiprovider Environment." ICCC99, 1999.

[92] ISO/IECJTC1/N10979, Working Draft for Open Distributed Processing Reference Model-Quality of Service. ISO, 1998.

[93] J. Asensio and J. Villagra, "A UML profile for QoS Management Information specification in distributed object based applications." Object Based Applications Proceedings, 1999.

[94] W. Stallings, High-Speed Networks and Internets: Performance and Quality of Service. Upper Saddle River, New Jersey: Prentice-Hall, Inc 2002.

[95] A. Jamalipour, Wireless Broadband Multimedia and IP Applications. Next Generation Wireless Networks. S. Tekinay. New Jersey, USA, Kluwer Academic Publishers: 264, 2000.

[96] J. Postel, User Datagram Protocol, available: http://www2.real-time.com/rte-ascend/1997/Nov/msg10389.html.

[97] J. Postel, Transmission Control Protocol, available: http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc0793.html.

[98] IETF-RFC1883, RFC 1883 - Internet Protocol, Version 6 (IPv6) Specification, available: http://www.faqs.org/rfcs/rfc1883.html.

[99] IETF-RFC1884, RFC 1884 - IP Version 6 Addressing Architecture, available: http://www.faqs.org/rfcs/rfc1884.html.

[100] IETF-RFC1885, RFC 1885 - Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6), available: http://www.faqs.org/rfcs/rfc1885.html.

[101] IETF-RFC1886, RFC 1883a - Internet Protocol, Version 6 (IPv6) Specification, available: http://www.faqs.org/rfcs/rfc1883a.html.

[102] IETF-RFC1970, RFC 1970 - Neighbor Discovery for IP Version 6 (IPv6), available: http://www.faqs.org/rfcs/rfc1970.html.

[103] IETF-RFC2133, RFC 2133 - Basic Socket Interface Extensions for IPv6, available: http://www.faqs.org/rfcs/rfc2133.html.

[104] 6.0 Quality of Service, available: http://www.ncs.gov/n2/content/tibs/html/tib97_2/sec6_0.htm.

[105] C. CIT, TCP protocols, available: http://nim.cit.cornell.edu/usr/share/man/info /en_US/a_doc_lib/aixbma. . . /tcp_protocols.html.

[106] G. Xylomenos and G. C. Polyzos, "Internet Protocol Performance Over Networks with Wireless Links." IEEE Network: 1999.

[107] J. Borland, State Looks to power companies for rural broadband. CNET News, 2000.

[108] T. Wallack, Long Wait for Speedy Service; Broadband Installation Can Drag. San Francisco Chronicle, 2000.

[109] history of satellite, available: http://celestrak.com/columns/v01n01.

[110] J. F. Graham, Chapter 18: The Space Station, available: www.space.edu/projects/book/chapter18.html.

[111] A. C. Clarke, "Extra Terrestrial Relays," Wireless World: pp. 305-308, October 1945 1945.

[112] L. A. Hogle, Satellite Communication, available: http://www.interlinx.qc.ca/leehogle/satellite.html.

[113] A. B. Carlson, Communication Systems an Introduction to Signals and Noise in Electrical Communication. singapore: McGraw-Hill, Inc. 1986.

[114] S. Adamson, B. Smith, D. Roberts, et al., Advance Satellite Communications: Potential Markets. Park Ridge, New Jersey, USA: Noyes Publications 1995.

[115] F. Ananasso and F. D. Priscoli, "Satellite systems for personal communication networks." Wireless networks 4: 1998, pp. 155-165.

[116] FAA and COMSTAC, 2003 Commercial Space Transportation Forecasts, May 2003, available: http://ast.faa.gov.

[117] A. Jamalipour, Low Earth Orbital Satellites for Personal Communication Networks. Norwood, MA: Artech House, Inc 1997.

[118] L. Wood, Satellite Constellation Networks. Chapter 2, Internetworking and Computing over Satellite Networks. Y. Z. (ed.). Doordrecht, Kluwer Academic Press,: pp. 13-34, 2003.

[119] P. v. Rossum, Gids Voor Satelliet Ontvangst. Deventer Antwerpen: Kluwer Technische boeken 1991.

[120] kepler, available: http://www.adsc.dial.pipex.com/kepler.htm.

[121] amsat, Keplerian Elements, 4 September 2003, available: http://www.amsat.org/amsat/keps/menu.html.

[122] telecable, The Keplerian Elements, available: http://www.telecable.es /personales/ea1bcu/kepsen.htm.

[123] L. Wood, Lloyds's Satellite Constellation, 6 September 2003, available: http://www.ee.surrey.ac.uk/Personal/L.Wood /constellations/background.html.

[124] H. Keller, H. Salzwedel, U. Freund, et al., Examination of The Circular Polar Satellite Constellation For The Use of Intersatellite Links,

[125] M. T. University, Map Projections, available: http://www.ice.mtu.edu/online_docs/TeraScan-3.2/html/terapgs/map_projections.html.

[126] Motorola, IRIDIUM, available: http://www.iridium.com/.

[127] Teledesic, Technical Overview of The Teledesic Network, available: http://www.teledesic.com/default.htm.

[128] A. S. Tanenbaum, Computer Networks. New Jersey: Prentice Hall 1996.

[129] L. Schiff and A. Chockalingam, "Signal Design and System Operation of Globalstar Versus IS-95 CDMA-Similarities and Differences." Wireless Networks 6: 2000, pp. 47-57.

[130] P. Chitre and F. Yegenoglu, "Next-Generation Satellite Networks: Architectures and Implementations," IEEE Communications Magazine, 37 (3): pp. 30-36, 1999.

[131] H. Kobayashi and B. L. Mark, "Generalized Loss Models and Queueing-Loss networks." IFORS'99, Beijing, China, 1999.

[132] N. Ansari, A. Arulambalam and S. Balasekar, "Traffic management of a satellite communication network using stochastic optimization." IEEE Transactions on neural network 7(3): 1996, pp. 732-744.

[133] P. Todorova, S. Olariu and H. N. Nguyen, "A Two-Cell-Lookahead Call Admission and Handoff Management Scheme for Multimedia LEO Satellite Networks." the 36th Hawaii International Conference on System Sciences (HICSS'03), Hawaii, IEEE Computer Society, 2002.

[134] D. P. Gerakoulis, W.-C. Chan and E. Geraniotis, "Throughput Evaluation of a Satellite-Switched CDMA (SS/CDMA) Demand Assignment System." IEEE Journal on Selected Areas in Communications 17(2): 1999, pp. 287-302.

[135] Y.-W. Chang and E. Geraniotis, "Optimal Policies for handoff and channel assignment in networks of LEO satellites using CDMA." Wireless Networks 4: 1998, pp. 181-187.

[136] G. Bianchi, N. Blefari-Melazzi, P. M. L. Chan, et al., "Design and Validation of QoS Aware Mobile Internet Access Procedures for Heterogeneous Networks." Mobile Networks and Applications 8: 2003, pp. 11-25.

[137] J. Sun and E. Modiano, "Routing Strategies for Maximizing Throughput in LEO Satellite Networks." IEEE Journal on Selected Areas in Communications: to be published in 2004.

[138] J. Sun and E. Modiano, "Capacity Provisioning and Failure Recovery for Low Earth Orbit Satellite Networks." International Journal on Satellite Communications 21: 2003, pp. 259-284.

[139] D. Corne, M. Dorigo and F. Glover, New Ideas in Optimization. Norwich UK: McGraw-Hill co. 1999.

[140] E. Aarts and J. K. Lenstra, Local Search in combinatorial optimization. New York: Chichester England 1997.

[141] P. E. Black, NP definition, available: http://hissa.nist.gov/dads/HTML/np.html.

[142] B. Dengiz, F. Altiparmak and A. E. Smith, "Local search genetic algorithm for optimal design of reliable networks." IEEE Transactions on Evolutionary Computation 1(3): September 1997, pp. 179-188.

[143] M. Mitchell, An Introduction to Genetic Algorithms. London, England: MIT Press, London England 1996.

[144] R. Sedgewick, Algorithms. Massachusets: Addison Wesley 1988.

[145] G. E. Miller, Technological Evolution as Self-fulfilling Prophecy: From Genetic Algorithms to Darwinian Engineering. Technological Evolution as self-fulfilling prophecy. J. Ziman. Cambridge, Cambridge University Press.: 203-215, 2000.

[146] D. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning: Addison-Wesley Publishing Company, Inc. 1989.

[147] G. J. Deboeck, Trading on The Edge : Neural, Genetic, and Fuzzy Systems For Chaotic Financial Markets. New York: Wiley, 1994.

[148] H. Chou, G. Premkumar and C. H. Chu, "Genetic Algorithms for communication network design-an empirical study of the factors that influence performance." IEEE Transactions on evolutionary computation 5(3): 2001, pp. 237-249.

[149] L. Berry, B. Murtagh, G. McMahon, et al., "Fast Network Design for Telecommunications." APORS2000, Singapore, 2000.

[150] G. McMahon, R. Septiawan, S. Sugden, "A Multiservice Traffic Allocation Model for LEO Satellite Communication Networks." IEEE Journal on Selected Areas in Communications 22(3): to be published in 2004.

[151] E. J. Montgomery, Crossover Mechanisms for Chromosomal Representation of Flow Representation of Flow Patterns in A Communication Network,

[152] L. T. M. Berry, B. A. Murtagh, G. McMahon, et al., "An Integrated GALP approach to communication network design." 2nd IFIP Workshop on traffic management and synthesis of ATM networks, Canada, 1997.

[153] W. J. Cook, W. H. Cunningham, W. R. Pulleyblank, et al., Combinatorial Optimization. New York US: John Wiley&Sons Inc. 1998.

[154] S. I. Gass, Linear Programming Methods and Applications. New York: Mc Graw-Hill Book Company 1975.

[155] B. A. Murtagh and S. J. Sugden, "A Direct Search Approach To Nonlinear Integer Programming." Optimization Methods and Software 4: 1994, pp. 171-189.

[156] F. Glover and M. Laguna, Tabu Search: Kluwer Academic Publishers 1997.

[157] J. Xu, S. Y. Chiu and F. Glover, Tabu Search for Dynamic Routing Communications Network Design, December 1996,

[158] J. Xu, S. Y. Chiu and F. Glover, Probabilistic Tabu Search for Telecommunications Network Design,

[159] T. H. Cormen, C. E. Leiserson and R. L. Rivest., Introduction to Algorithms. New York: Cambridge, Mass. : MIT Press ; New York : McGraw-Hill, 1990.

[160] T. H. o. C. Project, Edsger Wybe Dijkstra, November 7 2003, available: http://www.thocp.net/biographies/dijkstra_edsger.htm.

[161] T. S. Kelso, History of Satellite, available: http://celestrak.com/columns/v01n01.

[162] K. S. Meier Hellstern and W. Fischer, The Markov-Modulated Poisson Prosess (MMPP) cookbook 1992.

[163] E. J. Montgomery, Gennet, Development of a Genetic Algorithm for The Design of A Telecommunication Network. School of Information Technology. QLD, Bond University, 1999:pp 56.

[164] R. W. Wolff, Stochastic Modelling and the Theory of Queues. New Jersey: Prentice Hall, Englewood Cliffs, NJ 1989.

[165] B. Maglaris, D. Anastassiou, S. Prodip, et al., "Performance Models of Statistical Multiplexing in Packet Video Communications." IEEE Trans. On Communications 36(7): 1988, pp. 834-844.

[166] F. Brochin and J. Thomas, A three state Markov Chain Model For Speech Dynamics and Related Statistical Multiplexer Delay Performance. A three state Markov Chain Model For Speech Dynamics and Related Statistical Multiplexer Delay Performance. J. F. Inst., J. Franklin Inst. vol.327: 903-921, 1990.

[167] P. O'Reilly and S. Ghani, "Data performance in Burst Switching When The Voice Silence Periods Have A Hyperexponential Distribution." IEEE Trans. On Communication 35(10): 1987, pp. 1109-1112.